أداة بحث مقترحة لاسترجاع المعلومات في مجال مشاركة الملفات من نظير إلى نظير: دراسة تحليلية تجريبية

A proposed research tool for retrieving information in the peer-to-peer file sharing domain: a pilot analytical study

إعداد

د. محمد كامل احمد عبد الجواد

مدرس نظم استرجاع المعلومات - جامعة بني سويف– كلية الآداب

Doi:10.33850/ajahs.2021.140339

الاستلام: ٦/ ١١ /٢٠٢٠ القبول : ٢٣ / ١١ /٢٠٢٠

المستخلص :

تم تطوير أداة للبحث في استرجاع المعلومات في مجال مشاركة الملفات من نظير إلى نظير. تعتمد أداتنا "IR-P2P" على بروتوكول Gnutella الشهير، مما يتيح لنا الوصول إلى قاعدة مستخدمين كبيرة ومجموعة كبيرة من البيانات؛ حيث تحتفظ IR-P2P بالعديد من الإحصاءات وتنفذ عددًا من وظائف تصنيف ومعالجة واسترجاع المعلومات. تركيزنا الرئيسي هو انها أداة بحث، لذا يحتوي IR-P2P على مستودع بيانات ومحلل. يقوم مستودع البيانات بتخزين كل من الاستعلامات الواردة والصادرة ونتائج الاستعلام ويوفر طريقة لإنشاء صورة لكامل مجموعة البيانات التي شاركها المستخدمون. ويوفر محلل البيانات واجهة مستخدم بسيطة لتحليل البيانات. وسوف نناقش باختصار تحليلًا تم إجراؤه على مليون استفسار وارد تم جمعها في ملفات السجل الخاصة بالأداة المقترحة.

الكلمات المفتاحية: استرجاع المعلومات – شبكة نظير إلى نظير p2p - فهرسة البيانات – مرشح بلوم

Abstract:

A search tool was developed to retrieve information in peer-to-peer file sharing. Our "IR-P2P" tool is based on the famous Gnutella protocol, that allowing us to **access** a large user base and a wide range of data. As a research tool, IR-P2P maintains many statistics and implements a number of

information retrieval functions. Our main focus is that it is a search tool, so IR-P2P contains a data logger and analyzer. The data logger records both incoming and outgoing queries and query results and provides a way to create an image of the entire set of data shared by users. The Data Analyzer provides a simple user interface for data analysis. We briefly discuss an analysis one million incoming queries collected in the log files of the proposed tool.

المقدمة المنهجية

١ / ٠ تمهيد

في شبكة نظير إلى نظير، تكون كل عقدة متساوية ويمكنها توفير واستهلاك الموارد. تتكون شبكة نظير إلى نظير عادة من الآلاف من الآلات منخفضة التكلفة، وجميعها مزودة بقدرات معالجة وتخزين تتسم بالكفاءة بالإضافة إلى سرعات ارتباط عالية الكفاءة. تتمثل مزايا هذه الشبكات في عدم وجود نقطة مركزية للفشل والتنظيم الذاتي والتوسع الذاتي. في هذا البحث ركزنا على شبكات استرجاع المعلومات المستخدمة لتبادل البيانات العامة، ولقد حددنا عددًا من التحديات الرئيسية لشبكات الند للند فيما يتعلق بضمانات الاستخدام وسلوك الأقران والتقييم. قدمنا المهام الرئيسية الثلاث التي يؤديها كل نظام نظير إلى نظير: البحث وتحديد الموقع والنقل. علاوة على ذلك، قمنا بتنظيم الهياكل الرئيسية حول موضع المؤشر: الفهرس العمومي المركزي، الفهرس العمومي الموزع، الفهارس المحلية الصارمة والفهارس المحلية المجمعة. كل لها عواقب وخيمة لتوجيه الاستعلام والمعالجة. لقد أظهرنا أيضًا المسافة الفاصلة بين فهرس مؤلف من خطوة واحدة، حيث يتم تعيين الكلمات المفتاحية مباشرة إلى الملفات النصية، وفهرس من خطوتين، حيث يتم تعيين الكلمات المفتاحية إلى الأقران والأقران أنفسهم يكون لديهم تعيينات للملفات النصية.

في استرجاع المعلومات من نظير إلى نظير، تتمثل المهمة المركزية في البحث في تناقض مع نظم مشاركة الملفات حيث يكون التركيز على النقل. تتداخل بشكل ملموس العديد من التحديات في استرجاع المعلومات الموحدة مع تلك الموجودة في استرجاع معلومات نظير إلى نظير: وصف المورد واختيار المجموعة ودمج النتائج. ومع ذلك، في استرجاع المعلومات الموحدة، يلعب الطرف الوسيط دورًا مهمًا، ويظهر أيضًا تقسيمًا صارمًا بين المستفيدين ومقدمي المعلومات.

في البحوث الحالية لاسترجاع المعلومات من نظير إلى نظير، نجد كلا من منهجي فهرسة التقسيم و فهرسة التقسيم بكلمة رئيسية. لقد ظهرت العديد من التحسينات التي

يمكن تطبيقها لتقليل عرض النطاق الترددي ووقت الاستجابة ولتحسين جودة وكمية نتائج البحث التي يتم استرجاعها. إلى جانب هذا قدمنالمحة عامة وتصنيف للنظم الحالية. وأخيرًا، ناقشنا مستقبل استرجاع معلومات نظير إلى نظير، مما يشير إلى التحديات النوعية والمجالات الرئيسية التي يجب التركيز عليها. وأهمها: التأكيد على الدقة في الاسترجاع، والتركيز على نتائج البحث بدلاً من الملفات، والجمع بين نقاط القوة للمؤشرات المحلية والعمومية، وتطبيق التجميع واستخدام الملاحظات ذات الصلة.

٠ / ٢ مشكلة الدراسة

تحظى تقنية نظير إلى نظير (**P2P**) بشعبية كبيرة، حيث تعد مشاركة الملفات أحد تطبيقات **P2P** الرائدة. مع زوال **Napster** في يوليو ٢٠٠١ ، اكتسب كلا النظامين شبه المركزيين مثل **FastTrack** وأنظمة تبادل الملفات **P2P** اللامركزية تمامًا مثل **Gnutella** شعبية. **Gnutella** هي شبكة مشاركة الملفات الأكثر شعبية في الإنترنت. يُعتقد أنه يستضيف في المتوسط حوالي ٢,٢ مليون مستخدم يوميًا، على الرغم من أن حوالي ٤٠٠٠٠٠ إلى ٥٠٠٠٠٠ مستخدم عبر الإنترنت يكونوا فاعلين في أي لحظة معينة. بسبب الشعبية الهائلة لهذه الشبكات، ولذلك فلا بد من إيجاد طرق لتحسين أدائها. كما أن هناك حاجة إلى أدوات بحثية لتحويل استخدام هذه التقنيات إلى تجربة سهلة. ويساعد في ذلك دراسة كل من البيانات المتدفقة عبر هذه الشبكات وسلوك المستخدمين وتفضيلاتهم. على الرغم من وجود هذه الأدوات للأنظمة المركزية، إلا أننا لا نعلم بوجود نظائر **P2P** متوفرة بسهولة.

٠ / ٣ تساؤلات الدراسة: تعمل الدراسة على الإجابة على مجموعة من التساؤلات التالية :-

١. ما هي التقنيات المستخدمة في استرجاع المعلومات في مجال مشاركة الملفات في شبكات نظير إلى نظير؟
٢. ما هي النظم العملية المستخدمة في استرجاع المعلومات في مجال مشاركة الملفات في شبكات نظير إلى نظير؟
٣. كيف يمكن تطوير أداة بحث لاسترجاع المعلومات في مجال مشاركة الملفات في شبكات نظير إلى نظير؟
٤. ما هي فعالية أداة البحث المقترحة؟

٠ / ٤ أهداف الدراسة:

تسعى الدراسة إلى تحقيق مجموعة الأهداف التالية :-

١. التعرف على مختلف التقنيات المستخدمة في استرجاع المعلومات في مجال مشاركة الملفات في شبكات نظير إلى نظير.

٢. التعرف على الانظمة العملية المستخدمة في استرجاع المعلومات في مجال مشاركة الملفات في شبكات نظير إلى نظير.
٣. تطوير أداة بحث لاسترجاع المعلومات في مجال مشاركة الملفات في شبكات نظير إلى نظير.
٤. تقييم أداة البحث المقترحة.

٠ / ٥ أهمية الدراسة :

تزداد أهمية مفاهيم وتطبيقات شبكات نظير إلى نظير يوما بعد يوم، وفي ذات الوقت تتنوع التقنيات المختلفة المستخدمة في البحوث والنظم العملية لاسترجاع المعلومات في هذه الشبكات ومن هنا يساعد **IR-P2P** في إنشاء مجموعة استعلام غير متحيزة؛ حيث إنه يركز على أبحاث استرجاع المعلومات في تطبيق واسع النطاق قائم على **Gnutella**، ومشاركة الملفات، مما يسمح للمستخدمين بجمع المزيد من الاستعلامات الحديثة وتوفير واجهة سهلة لإجراء التحليل على مجموعة البيانات.

٠/ ٦حدود الدراسة

الحدود الموضوعية: عملية التقييم الأداة المقترحة ستقتصر على الجوانب المحددة بالبحث الحالي.

الحدود النوعية: تقتصر أداة البحث المقترحة على مجال مشاركة الملفات في شبكات نظير إلى نظير ولا تنسحب على باقي شبكات مشاركة الملفات ولا باقي محركات البحث.

٠ / ٧ منهجية الدراسة

اعتمد الباحث على **المنهج التجريبي** في الدراسة الحالية باعتباره المنهج المناسب لطبيعة موضوع الدراسة والأكثر استخداما في مجال مشاركة الملفات من نظير إلى نظير؛ حيث قام الباحث بتصميم أداة بحث لاسترجاع المعلومات في شبكات مشاركة الملفات من نظير إلى نظير وتمت تجربة الأداة المقترحة للحكم على فعاليتها وكفاءة استرجاعها للمعلومات، هذا بالإضافة إلى أن هذا المنهج يساعد الباحث على التحكم في المتغيرات المختلفة التي قد تؤثر على نتائج التقييم مثل اختلاف خلفية الباحث وموضوع البحث ومهارات الباحث في عملية البحث ذاتها.

وبجانب المنهج التجريبي تم الاعتماد على **المنهج الوصفي التحليلي** الذي يعمل على بتحليل نتائج عمليات البحث المختلفة في نظم استرجاع المعلومات المبنية على مشاركة الملفات من نظير الى نظير.

أما **أدوات البحث** فتتمثل في أداة البحث الوثائقي في جمع الانتاج الفكري المرتبط بظاهرة كمشاركة الملفات من نظير الى نظير، واستخدام أداة الابحار والمعايشة مع الويب الى جانب استخدام أداة الملاحظة في الجانب التطبيقي للبحث.

عينة الدراسة :

تم تقييم اداة البحث المصممة من خلال ٧٧٥٦٠٥ استعلام تم عن طريق أداة البحث المقترحة وتنوعت الاستعلامات ما بين طلب المستندات والملفات الصوتية والمرئية.

٠ / ٨ مصطلحات الدراسة

- **شبكة نظير إلى نظير:** شبكة نظير إلى نظير **(P2P)** هي مجموعة من أجهزة الكمبيوتر، كل منها يعمل كعقدة لمشاركة الملفات داخل المجموعة. بدلاً من وجود خادم مركزي للعمل كمحرك مشترك، يعمل كل كمبيوتر كخادم للملفات المخزنة عليه. عند إنشاء شبكة نظير إلى نظير **(P2P)** عبر الإنترنت، يمكن استخدام خادم مركزي لفهرسة الملفات، أو يمكن إنشاء شبكة موزعة حيث يتم تقسيم مشاركة الملفات بين جميع المستخدمين في الشبكة التي تخزن ملفًا معينًا.
- **نظم استرجاع الويب :** يمثل نظام الاسترجاع كيان رقمي يعمل في بيئة الويب لمعالجة وبحث مواد الويب واتاحتها وفقا لطبيعة ونوعية الحاجات والمجالات الموضوعية والتوجهات الفكرية مسار بحث المستخدم.

٠ / ٩ الدراسات السابقة

اعتمد الباحث في البحث عن الدراسات المثيلة والسابقة التي اهتمت بدراسة ظاهرة مشاركة البيانات والملفات من نظير إلى نظير على قواعد البيانات العالمية المتضمنة بيانات اتحاد المكتبات الجامعية المصرية **EULC** مثل **springer, science direct, IEEE, Dissertation proquest** وغيرها. وقد نتج عن هذا البحث العديد من الدراسات التي حول نظم نظير إلى نظير التي تركز على على كيفية تصميم شبكة نظير إلى نظير. ويتعامل معظمهم مع نمذجة المستويات الدنيا لشبكة نظير إلى نظير، مثل نمذجة بروتوكولات النظراء، أو نمذجة الطريقة التي يتم بها نسخ الملفات على الشبكة، أو نمذجة طبولوجيا الشبكة[١]. لا يركز عملنا على المستوى الأدنى للشبكة، ولكن على مستوى التطبيق لنظام استرجاع معلومات نظير إلى نظير.

هناك عدد من مشاريع استرجاع المعلومات من نظير إلى نظير التي تم تطويرها، مثل **Peerware** و **Anthill** و **Alvis** و **Nutch**، على سبيل المثال لا الحصر. تركز معظم هذه المشروعات على توجيه الاستعلام، أو اكتشاف المورد، أو بنية تصميم نظام استرجاع معلومات نظير إلى نظير. معظمهم يطبقون إمكانات البحث، لكن لديهم تركيز أقل على تسجيل البيانات وتحليلها٢.

تقوم معظم الأعمال الحالية على تحليل سجل الاستعلام لمحركات البحث على الويب. سيلفرشتاين وآخرون (١٩٩٩)٣ في دراستهم " Analysis of a Very Large Web Search Engine Query Log" قاموا بتحليل مجموعة كبيرة جدًا من طلبات البحث التي سجلها AltaVista لمدة ستة أسابيع، ويحتوي على ما يقرب من مليار استفسار.

قام بيتزل وآخرون (٢٠٠٤)٤ في دراستهم " Hourly Analysis of a Very Large Topically Categorized Web Query Log" بتحليل تغييرات الاستعلامات من حيث شعبية الاستعلام والتفرد مع مرور الوقت. وقام زينالبور - يازتي وفولياس (٢٠٠٢) في دراستهم " A Quantitative Analysis of the Gnutella Network Traffic" بتسجيل رسائل الاستعلام وتحليلها لشبكة Gnutella في عام ٢٠٠٢. من خلال جمع استفسارات المستخدم التي تم نشرها. الأداة الخاصة بهم قامت ١٧ محطة عمل لجمع كل رسائل المرور. في غضون ٥ ساعات، قاموا بجمع حوالي ١٥ مليون رسالة استعلام. المحددات من النهج الذي اتبعوه هو أنه لا يتم توجيه كل رسائل الاستعلام إلى أقرانهم. على الأرجح، تتشابه رسائل الاستعلام التي تلقوها مع ما كانوا يشاركونه. وبالتالي، قد لا تمثل مجموعة الاستعلام الخاصة بها جميع طلبات البحث التي تتدفق عبر الشبكة. في تطويرنا لأداة البحث IR-P2P قمنا بتعديل جداول التوجيه الخاصة بنا لالتقاط جميع الاستعلامات التي تتدفق عبر الشبكة. لذلك، نتلقى مجموعة كاملة من جميع استعلامات المستخدم في فترة زمنية محددة. بالإضافة إلى ذلك، جمع زينالبور - يازتي وفولياس استعلامات نظير إلى نظير في يونيو ٢٠٠٢. لذلك، قد تكون مجموعة طلبات البحث قديمة، لأن ما يبحث عنه المستخدمون منذ ١٥ سنة قد يكون مختلفًا تمامًا عما يبحثون عنه اليوم٥.

المبحث الأول: استرجاع المعلومات في شبكات نظير إلى نظير

١ / ١ مقدمة

في شبكة لاسترجاع المعلومات من نظير إلى نظير، فإن المهمة المركزية هي البحث: في حالة تقديم استعلام، يتم إسترجاع بعض قائمة مراجع الملفات: نتائج البحث. يمكن أن ينشأ الاستعلام عن أي نظير في الشبكة، ويجب أن يتم توجيهه إلى واحد أو أكثر من الأقران الآخرين الذين يمكنهم تقديم نتائج البحث على أساس الفهرسة. ويمكن للاقران استعراض وتقديم النتائج. نتيجة البحث عبارة عن تمثيل مضغوط للملف. يمكن أن يحتوي الملف على نص أو صورة أو صوت أو فيديو أو مزيج من هذه الوسائط. تسمى نتيجة البحث أحيانًا مقتطفًا وتتضمن على الأقل فهرس للوثيقة الكاملة والبيانات الوصفية الإضافية الشائعة مثل العنوان وملخص وحجم الملف، إلخ. ترتبط

كل نتيجة معروضة بالملف الكامل المرتبط. يوفر التمثيل المضغوط فرصة أولية للمستخدمين لتمكينهم من اختيار الروابط التي يريدون متابعتها٦.

يمكن تقسيم شبكات استرجاع معلومات نظير إلى نظير إلى فئتين بناءً على موقع الملفات المشار إليها. أولاً، المراجع ذات الملف الداخلي حيث يتعين تنزيل الملفات من أقرانهم الآخرين داخل الشبكة، على سبيل المثال: المكتبات الرقمية. ثانياً، أولئك الذين لديهم مراجع خارجية للوثائق حيث يتم الحصول على الملفات الفعلية وتحديد موقعها ونقلها خارج نطاق شبكة نظير إلى نظي ، على سبيل المثال: محرك بحث على الويب من نظير إلى نظير. في الأقسام التالية، نقوم بمقارنة شبكات استرجاع معلومات نظير إلى نظير مع التطبيقات والنماذج الأخرى[٧].

١ / ٢ مقارنة مع شبكات تبادل الملفات

تُستخدم شبكات مشاركة الملفات للبحث عن الملفات التي يتشاركها مستخدمي شبكة نظير إلى نظير وتحديد موقعها واسترجاعها. يشبه البحث في مثل هذه الشبكات استرداد معلومات نظير إلى نظير. يتم إدخال استعلام نصي يتم بعده إرجاع قائمة الملفات. بعد البحث، يقوم المستخدم باختيار نوع من الاهتمام للتنزيل والذي عادةً ما يحتوي على نوع من المعرّف الفريد، مثل التجزئة القائمة على المحتوى. والخطوة التالية هي تحديد موقع أقرانهم الذين لديهم نسخة من الملف. يمكن بعد ذلك نقل الملف من نظير معين، أو من عدة أقران في وقت واحد وفي هذه الحالة، يتم طلب أجزاء معينة من الملف من كل نظير يتم تجميعه مرة أخرى بعد اكتمال التنزيلات٨.

إن مهام تحديد موقع الأقران وخاصة نقل المحتوى هي التطبيق الأساسي لشبكات المشاركة للملفات ومجال التركيز على تحسينات البحث والأداء. عمليات البحث في هذه الشبكات عادة ما تكون للعناصر المعروفة، بينما في شبكات استرجاع المعلومات يكون القصد أكثر تنوعًا. بينما توفر بعض شبكات استرجاع المعلومات أيضًا عمليات تحديد المواقع وتنزيلها، إلا أنها عادةً ما تركز على تحسينات لمهمة البحث. إلى جانب هذا الاختلاف العام في التركيز، هناك ثلاثة اختلافات ملموسة على الأقل أيضًا٩.

أولاً، يعتمد فهرس البحث عن مشاركة الملفات عادةً فقط على أسماء الملفات المتوفرة على الشبكة وليس على محتواها كما هو الحال بالنسبة لاسترجاع المعلومات. فهرس الأسماء هذا أصغر من فهرس الملف الكامل. وبالتالي، هناك أيضًا عدد أقل من عمليات النشر لكل فصل مما يجعل تكلفة تقاطعات قوائم النشر أقل تكلفة، وهي عملية شائعة في أي فهرس عمومي موزع. نظرًا لصغر حجمها، يتم احتساب فهرس مركزي جيد لهذه الفهارس المستندة إلى الاسم. ومع ذلك، فقد أصبحت الشبكات التي يتم فحصها مركزيًا غير مرغوب فيها إلى حد كبير لأسباب قانونية.

ثانياً، عند إضافة ملف إلى فهرس مشاركة الملفات، لا يتغير. إذا كانت هناك حاجة إلى نسخة معدلة، تتم إضافتها ببساطة كقناة جديدة. وبالتالي، تحديثات الفهرس غير

مطلوبة. على النقيض من ذلك في شبكة استرجاع المعلومات عندما تتغير الوثيقة الأساسية، يجب تغيير نتائج البحث المرتبطة التي تم إنشاؤها من هذا الملف. وبالتالي، يجب تحديث الفهرس بحيث تعكس نتائج البحث التغييرات التي تمت على الملف المشار إليه.

ثالثًا، نظرًا لأن التركيز في شبكة مشاركة الملفات ينصب على تنزيل البيانات بأسرع وقت ممكن، فمن المهم أن تتمتع بإنتاجية عالية. في المقابل، في عملية استرجاع المعلومات، تهيمن مهمة البحث التي يكون فيها زمن الاستجابة المنخفض هو الأكثر أهمية. بشكل أكثر تحديداً: فمن المقبول إذا استغرقت الشبكة نصف دقيقة لتحديد موقع أسرع أقرانهم للتنزيل منه، في حين أن قضاء هذا الوقت غير مقبول للحصول على نتائج بحث جيدة. بالنسبة لمشاركة الملفات، فإن الفهرس المعروض هو الفهرس المستخدم لتحديد موقع الملف المحدد، في حين أنه بالنسبة لاسترجاع المعلومات فهو للبحث. يتم إجراء تعيين وهي الخطوة الأولى دائمًا على مستوى شبكة نظير إلى نظير بأكملها، بينما يتم إجراء تعيين كخطوة ثانية على نظير محدد.

١ /٣ مقارنة مع استرجاع المعلومات الموحدة.

في استرجاع المعلومات الموحدة، هناك ثلاثة أطراف: العملاء الذين يطرحون استعلامات ووسيطًا واحدًا ومجموعة من خوادم البحث التي يكشف كل منها عن مجموعة من الملفات: تشبه الفهارس المحلية الصارمة. تبدأ عملية البحث عندما يصدر عميل استعلامًا للوسيط. ويمتلك الوسيط معرفة بعدد كبير من خوادم البحث وجهات الاتصال التي تحتوي على مجموعة فرعية مناسبة منها للإجابة على الاستعلام. ثم يعرض كل خادم بحث مجموعة من نتائج البحث للاستعلام. يقوم الوسيط بدمج هذه النتائج في قائمة واحدة وإرجاعها إلى العميل.

وهناك ثلاثة تحديات تشكل أعمدة استرجاع المعلومات المتحدة التي تشترك فيها مع استرجاع المعلومات من نظير إلى نظير١٠. **أولاً،** مشكلة وصف المورد: يجب أن يتلقى الوسيط من كل خادم بحث إشارة إلى الاستعلامات التي يمكنه التعامل معها، في حالة الخوادم التعاونية، أو الوسيط بحاجة إلى معرفة ذلك عن طريق البحث في خوادم البحث إذا كانت غير متعاونة. في كلتا الحالتين تكون النتيجة النهائية هي وصف مورد لخادم البحث. عادةً ما تظل هذه الأوصاف صغيرة لأسباب تتعلق بالكفاءة، ويمكن وصف المجموعات الكبيرة بكمية صغيرة نسبيًا من البيانات. يمكن أن يتكون الوصف من، على سبيل المثال: إحصائيات موجزة، تقديرات حجم المجموعة، و / أو نموذج مستند تمثيلي. في شبكة استرجاع المعلومات من نظير إلى نظير، يحتاج الأقران إلى معرفة من هم أقرانهم الآخرون الذين يمكنهم إرسال استعلام. وبالتالي، هناك حاجة أيضا إلى أوصاف الموارد. تتمثل ميزة شبكات نظير إلى نظير في أن

الزملاء يتعاونون دائمًا ويتحدثون بروتوكولًا مصممًا ومتفقًا عليه، مما يجعل تبادل أوصاف الموارد أسهل إلى حد ما. ومع ذلك، قد يكون لدى الأقران حافز للغش في محتواها، مما يخلق تحديات فريدة خاصة بشبكات الند للند.

ثانياً، مشكلة اختيار المجموعة: بعد الحصول على أوصاف المورد، فإن الخطوة التالية هي اختيار مجموعة فرعية من خوادم البحث التي يمكنها معالجة الاستعلام. عندما يتلقى الوسيط استعلامًا جديدًا من عميل، يمكنه تسجيله محليًا بسرعة مقابل أوصاف المورد المكتسبة لتحديد الخوادم التي من المرجح أن تسفر عن نتيجة بحث ذات صلة بالاستعلام. يمكن تقسيم الخوارزميات الخاصة بتحديد أفضل الخوادم في استرجاع المعلومات الموحدة إلى مجموعتين. أولاً، تلك التي تتعامل مع وصف الموارد كمستندات كبيرة دون النظر في الملفات الفردية داخل كل مورد مثل: CVV ((((Card Verification Value و Kullback–Leibler divergence. ثانياً، تلك التي تنظر في المستندات الفردية داخل كل مورد. على الرغم من أن النظر في الوثائق الفردية يعطي نتائج أفضل، إلا أنه يزيد من تعقيد وصف الموارد وتكاليف الاتصال. بالإضافة إلى ذلك، تم تصميم معظم خوارزميات اختيار الموارد الحالية للاستخدام من قبل طرف وسيط واحد مما يجعلها غير صالحة للتطبيق في شبكة ذات فهارس محلية مجمعة، على سبيل المثال. يتطلب اختيار الموارد وفقًا للخصائص الفريدة لشبكات نظير إلى نظير تطوير خوارزميات جديدة.

ثالثًا، مشكلة دمج النتائج: بمجرد حصول الوسيط على نتائج من عدة خوادم بحث، يجب دمجها في قائمة واحدة متماسكة. إذا كانت جميع الخوادم تستخدم نفس الخوارزمية لترتيب نتائجها، فسيكون ذلك سهلاً. ومع ذلك، نادراً ما يكون هذا هو الحال ولا يتم تضمين درجات الترتيب الدقيقة بشكل شائع. تتمثل الخطوة الأولى في الدمج في تطبيع الدرجات على المستوى الأعلى، بحيث تكون مستقلة عن الموارد. في استرجاع المعلومات المتحدة، يمكن استخدام خوارزمية دمج التعلم تحت الإشراف (Secure Sockets Layer) (SSL) لهذا. ومع ذلك، في بيئات نظير إلى نظير، غالبًا ما تختلف مجموعات الملفات المفهرسة على نطاق واسع في أحجامها مما يجعل من المحتمل ألا يعمل SSL جيدًا. يتطلب SSL نموذج قاعدة بيانات يجعلها غير مرغوب فيها في شبكات نظير إلى نظير حذرة بشأن استخدام عرض النطاق الترددي. هناك طريقة بديلة تتمثل في إعادة حساب نتائج المستندات لدى الوسيط كما تفعل خوارزمية١١ **Kirsch 1997** وهي دقيقة تمامًا ولها تكاليف اتصال منخفضة من خلال مطالبة كل مورد بتوفير إحصائيات موجزة فقط. ومع ذلك، فإن هذا يتطلب أيضًا معرفة إحصائيات مجموعة القوانين العالمية التي يعد الحصول عليها مكلفًا في شبكات نظير إلى نظير ذات الفهارس المحلية. الخلاصة، يتطلب الدمج في شبكات استرداد معلومات نظير إلى نظير خوارزمية يمكنها أن تعمل بفعالية مع الحد الأدنى

من تكاليف التدريب الإضافية وتكاليف الاتصالات، والتي لا تتأهل مباشرة أي من خوارزميات دمج النتائج الحالية. لذلك اعتمدت عمليات الدمج في الشبكات الحالية حتى الآن على أساليب بسيطة تعتمد على التردد، ولم تقدم أي حل لتكامل النتائج القائم على الصلة.

لكن توجد عدة اختلافات وهي١٢، الاختلاف الأول الملحوظ في استرجاع المعلومات من نظير إلى نظير هو التخصص الدقيق للأطراف المختلفة. يصدر العملاء طلبات البحث فقط بينما تخدم خوادم البحث نتائج البحث فقط. يحدد هذا أيضًا شكل شبكة التراكيب الجامدة التي تتشكل: رسم بياني ثنائي الأطراف مع العملاء على جانب واحد، والخوادم على الجانب الآخر والوسيط في الوسط. في الواقع، يعد استرجاع المعلومات المتحدة أقرب بكثير من نموذج خادم العميل التقليدي ويشتمل عادةً على أجهزة "تعرف" بعضها البعض بالفعل. وهذا يتناقض مع شبكات نظير إلى نظير حيث يتولى الأقران هذه الأدوار حسب الحاجة ويتفاعلون في كثير من الأحيان بشكل فضفاض مع الأجهزة الأخرى "المجهولة". بالإضافة إلى ذلك، تخضع شبكة نظير إلى نظير لمشكلات كبيرة في التوافر وعدم التجانس والتي تؤثر فقط بشكل خفيف على شبكات استرجاع المعلومات المتحدة بسبب الفصل الصارم بين المخاوف.

الاختلاف الثاني هو وجود الوسيط. بالنسبة إلى العملاء، يظهر الوسيط كنقطة إدخال واحدة ويشكل واجهة: لا يدرك العملاء أبدًا وجود خوادم بحث متعددة على الإطلاق. وهذا ينطوي على أن جميع الاتصالات يتم توجيهها عبر الوسيط مما يجعلها نقطة اتصال واحدة. في الممارسة العملية، يمكن أن يكون الوسيط مزرعة خوادم لتخفيف ذلك. ومع ذلك، لا تزال هناك نقطة تحكم واحدة، على غرار أنظمة البحث المركزية تمامًا والتي يمكن أن تخلق صعوبات قانونية وأخلاقية. شبكة نظير إلى نظير مع نقطة "وسيط" مركزية لاستعلامات التوجيه قريبة من الناحية النظرية من شبكة استرداد المعلومات المتحدة. ومع ذلك، تميل معظم شبكات نظير إلى نظير إلى توزيع مهمة الوساطة هذه، وتعيين الاستعلامات إلى الزملاء الذين يمكن أن يوفروا نتائج بحث ذات صلة على نظيرات متعددة.

١ /٤ تقييم انظمة استرجاع المعلومات في شبكات نظير إلى نظير

إن النظرة العملية لهدف استرجاع المعلومات من نظير إلى نظير تقلل من عدد الرسائل المرسلة لكل استعلام مع الحفاظ على الاسترجاع والدقة العالية. هناك العديد من الأساليب التي تمثل بدائل مختلفة. دعنا نبدأ باستراتيجيتين شائعتين لتقسيم الفهارس على أجهزة متعددة: تقسيم وفق الملف وتقسيم وفق الكلمات الرئيسية. في التقسيم وفق الملف - كل نظير مسؤول عن الحفاظ على فهرس محلي يحتوي مجموعة محددة من الوثائق: توجد جميع منشورات وشروط وثيقة معينة في نظير

واحد محدد. في بعض الحالات ، يتم تخزين الملفات نفسها أيضًا على هذا النظير. تستخدم الفهارس المحلية الصارمة ومعماريات الفهارس المحلية بشكل شائع في شبكات نظير إلى نظير التي تستخدم هذا النوع من التقسيم. في المقابل، في التقسيم وفق الكلمات الرئيسية، يكون كل نظير مسؤولاً عن تخزين المنشورات لبعض الكلمات الرئيسية المحددة في الفهرس. البنية المعتادة لهذا هو الفهرس العمومي الموزع.

تم إجراء تحقيق مبكر في جدوى شبكة بحث الويب من نظير إلى نظير بواسطة لي وآخرون (٢٠٠٣)١٣. إنهم ينظرون إلى التقسيم وفق الملف كنقطة بداية أكثر قابلية للتتبع، لكنهم يوضحون أن التقسيم وفق الملف يمكن أن يدخل في نطاق أداء التقسيم وفق الكلمات الرئيسية عن طريق تطبيق تحسينات مختلفة على الفهرس العمومي الموزع. في المقابل استنتج سيول وآخرون (٢٠٠٣)١٤ أن مقاربة التقسيم تلائم المقياس بشكل سيئ، لأن مجموعات الملفات لا تجمع "بشكل طبيعي" بطريقة تسمح بتوجيه الاستعلام إلى جزء صغير من أقرانهم وبالتالي يتطلب كل استعلام الاتصال بجميع أقرانهم تقريبًا في النظام. ربما نظرًا لهذه الورقة، ركزت الكثير من الأبحاث في استرجاع المعلومات من نظير إلى نظير على التقسيم وفق الكلمات الرئيسية باستخدام الفهرس العمومي الموزع.

لسوء الحظ، لا يخلو الفهرس العمومي الموزع من عيوب لأنه يهدف إلى إجراء عمليات تحديد فعالة، وليس للبحث الفعال. أولاً، يوفر جدول التجزئة موازنة التحميل بسذاجة إلى حد ما، من خلال اللجوء إلى توحيد دالة التجزئة المستخدمة. نظرًا لاختلاف حجم قوائم نشر المصطلحات، فقد يتسبب ذلك في ظهور نقاط اتصال سريعة للمصطلحات الشائعة التي تؤدي إلى إلغاء توازن الحمل. ثانياً، إن تقاطع قوائم نشر المصطلحات المستخدمة في الفهارس العمومية الموزعة يتجاهل الارتباط بين المصطلحات التي يمكن أن تؤدي إلى دقة بحث غير مرضية. ثالثًا، تزداد تكلفة الاتصال الخاصة بالتقاطع بالتناسب مع عدد مصطلحات الاستعلام وطول القوائم المقلوبة. لقد تم اقتراح العديد من التحسينات مثل تخزين نتائج الاستعلام متعدد المدة لمدة محددة محليًا لتجنب التقاطعات ومطالبة كل نظير بتخزين معلومات إضافية للمصطلحات المرتبطة بشدة بالمصطلحات التي يخزنها بالفعل. وبالتالي، فإن اختيار أوصاف الموارد في فهرس عمومي موزع يتم تقييده بسبب ارتفاع تكاليف الاتصالات الخاصة بتحديثات الفهرس: من غير المرجح أن يعمل تمثيل النصوص الكاملة بشكل جيد بسبب حركة مرور الشبكة الهائلة التي يتطلبها ذلك. رابعًا، قد تؤدي إحصائيات المجموعة المنحرفة كنتيجة لتقسيم المدة إلى نتائج تصنيف لا تضاهى الفهرس العمومي الموزع. أخيرًا، تكون جداول التجزئة الموزعة عرضة لمجموعة متنوعة من هجمات الشبكات التي تهدد أمن وخصوصية المستخدمين١٥.

يفشل العديد من الباحثين في رؤية عدد من الفوائد الفريدة للفهارس المحلية لكل قسم على حدة، مثل التكاليف المنخفضة للعثور على العناصر الشائعة ومعالجة الاستعلام المتقدمة وتحديثات الفهرس الرخيصة والمقاومة المرتفعة للمضخات. من المسلم به أن التحدي الأساسي لهذه الفهارس هو توجيه الاستعلام إلى أقرانه المناسبين. موقفنا هو أن كلا النهجين لهما ميزة ويكمل كل منهما الآخر. تؤكد الأبحاث الحديثة بالفعل فاعلية استخدام الفهارس المحلية لمصطلحات الاستعلام الشائعة وفهرس عمومي لمصطلحات الاستعلام النادرة.

أن البحث على شبكة الإنترنت غير ممكن باستخدام تقنية نظير إلى نظير. تعتبر النفقات العامة التي يتم تقديمها عن طريق الاتصال بين الأقران أكبر من أن توفر أوقات استجابة معقولة للاستعلام نظرًا لقدرة الإنترنت. ومع ذل ، فقد تم إنجاز الكثير من العمل، الذي تمت مناقشته في القسم التالي، منذ الابحاث الاولى في هذا المجال وتغيرت طبيعة الإنترنت وقدرتها بشكل كبير خلال هذه الفترة.

قارن يانغ وآخرون (٢٠٠٦)١٦ قارنوا أداء العديد من بنيات نظير إلى نظير لاسترجاع المعلومات مع التحسينات الشائعة. يقومون باختبار ثلاثة طرق: فهرس عالمي موزع معزز بمرشح بلوم وذاكرة التخزين المؤقت؛ فهارس محلية مجمعة مع فيضان الاستعلام؛ وفهارس محلية صارمة باستخدام مسارات عشوائية. كل هذه هي فهارس وثيقة المدى من خطوة واحدة. ومن المثير للاهتمام، أنها تستهلك جميعها تقريبًا نفس مقدار عرض النطاق الترددي أثناء معالجة الاستعلام، على الرغم من أن الفهارس المحلية المجمعة هي الأكثر كفاءة. ومع ذلك، فإن الفهرس العمومي الموزع يقدم أقل زمن وصول لهذه الأساليب الثلاثة، يليه عن كثب فهارس محلية مجمعة وفهارس محلية صارمة تكون أوامرها أبطأ. بالنسبة لجميع الطرق، يقدم إعادة توجيه الاستعلامات في الشبكة أقصى امان، في حين أن الإجابة على الاستعلامات رخيصة نسبيًا. على الرغم من أن الفهرس العمومي الموزع سريع، إلا أن عيبه الرئيسي يظهر بوضوح في وقت الفهرسة والنشر. عند إضافة مستندات جديدة إلى الشبكة، يستخدم هذا النطاق الترددي ستة أضعاف الوقت مقارنة بالفهارس المحلية المجمعة لتحديث قوائم النشر. تعمل الفهارس المحلية الصارمة على حل كل هذا محليًا ولا تتحمل أية تكاليف من حيث الوقت أو النطاق الترددي لنشر المستندات. توضح هذه الدراسة بوضوح أن الهيكل يجب أن يحقق التوازن بين سرعة الاسترجاع وتكرار التحديث.

المبحث الثاني: تقنيات استرجاع المعلومات في شبكات نظير إلى نظير

في هذا القسم نناقش العديد من تقنيات تحسين استرجاع المعلومات في شبكات من نظير إلى نظير. هناك سببان لاستخدام هذه التقنيات. أحدهما هو تقليل استخدام عرض النطاق الترددي، والآخر هو تحسين جودة وكمية نتائج البحث التي يتم إسترجاعها.

تؤثر معظم التقنيات التي ستتم مناقشتها على كل من هذه الجوانب وتقدم مقايضات، على سبيل المثال: يمكن للمرء التضحية بالكمية لتوفير عرض النطاق الترددي وبالجودة لتقليل زمن الوصول.

٢ /١ التقاطع التقريبي لقوائم النشر مع مرشح بلوم والتباديل المستقلة للتجزئة الحساسة (MinHash)

عند استخدام فهرس عمومي موزع، يتطلب الاستعلام متعدد المدة عمليات بحث متعددة في جدول التجزئة الموزعة. يجب أن تتقاطع قائمة النشر لكل مصطلح للعثور على الوثائق التي تحتوي على جميع شروط الاستعلام. يمكن أن يكون تبادل قوائم النشر مكلفًا من حيث النطاق الترددي، خاصةً للمصطلحات الشائعة مع العديد من المنشورات، وبالتالي يمكن نقل مرشح بلوم الأصغر المشتق من هذه القوائم بدلاً من ذلك. تم استخدام مرشح بلوم لأول مرة في سياق استرداد معلومات نظير إلى نظير بواسطة رينولدز وآخرون (٢٠٠٣)١٧.

مرشح بلوم هو مجموعة من البتات. يتم تعيين كل بت في البداية إلى صفر. يمكن إجراء عمليتين على مرشح بلوم: إدخال قيمة جديدة واختبار ما إذا كانت القيمة الحالية موجودة بالفعل في المرشح. في كلتا الحالتين، يتم تطبيق دالات التجزئة الخطية **k** أولاً على القيمة. تقوم عملية الإدراج، بناءً على النتيجة، بتعيين المواضع **k** لمرشح بلوم إلى واحد. تقرأ اختبارات العضوية المواضع **k** من مرشح بلوم. إذا كان كل منهم يساوي واحدًا، فقد تكون هذه هي القيمة في مجموعة البيانات. ومع ذلك، إذا كان أحد المراكز **k** يساوي الصفر، فليس من المؤكد أن القيمة ليست في مجموعة البيانات. وبالتالي، النتائج الايجابية الكاذبة ممكنة، لكن النتائج السلبية الخاطئة لا تحدث أبداً. تعتبر مرشحات بلوم طريقة جذابة للبيئات الموزعة لأنها تحقق رسائل أصغر مما يؤدي إلى توفير كبير في الشبكة١٨.

المثال التالي في سياق استرجاع معلومات نظير إلى نظير يوضح ما سبق: النظير **Q** يطرح استعلام **q** يتكون من المصطلحات **a** و **b**. نحن نفترض أن المصطلح **a** لديه أطول قائمة نشر. يحتفظ النظير **A** بالمنشورات **Pa** للمصطلح **a**، ويستمد مرشح بلوم **Fa** من هذا ويرسله إلى النظير **B** الذي يحتوي على المنشورات **Pb** للمصطلح **b**. يمكن للنظير **B** الآن اختبار عضوية كل وثيقة في **Pb** مقابل مرشح بلوم **Fa** وإعادة إرسال القائمة المتقاطعة **Fa**∩**(P** إلى النظير **Q** كنتيجة نهائية. نظرًا لأن هذا قد لا يزال يحتوي على إيجابيات كاذبة، فيمكن بدلاً من ذلك إرسال التقاطع إلى النظير **A**، والذي يمكن أن يزيل الإيجابيات الخاطئة لأنه يحتوي على منشورات كاملة **Pa**، والنتيجة هي **Pa** ∩ **Pb** ∩ **Fa**: التقاطع الحقيقي للمصطلحين **a** و **b** والذي يمكن إرساله كنتيجة للنظير **Q**. تحدث وفورات النطاق الترددي عند إرسال

Fa الصغير بدلاً من **Pa** الكبير من النظير **A** إلى **B**. ومع ذلك يتطلب هذا النهج خطوة إضافية إذا كان أحد يريد إزالة النتائج الايجابية الكاذبة.
النتائج الايجابية الكاذبة هي أكبر عيب لمرشحات بلوم: فكلما قل عدد البتات المستخدمة، زاد احتمال حدوث خطأ كاذب. تتطلب المجموعات الكبيرة أن يتم تمثيل عدد أكبر من وحدات البت أكثر من المجموعات الأصغر. لسوء الحظ، تحتاج مرشحات بلوم إلى أن يكون لها نفس الحجم لعمليات التقاطع والاتحاد. هذا يجعلها غير مناسبة تمامًا لشبكات نظير إلى نظير التي يكون لدى أقرانها مجموعات تختلف اختلافًا كبيرًا في عدد المستندات التي تحتوي عليها.
يمكن استخدام مرشحات بلوم لإجراء التقاطع التقريبي لقوائم النشر. ومع ذلك، كخطوة قبل ذلك، من المثير للاهتمام أيضًا تقدير ما ستفعله قائمة النشر الإضافية من حيث التقاطع مع القوائم التي تم الحصول عليها بالفعل. هذه المهمة لا تتطلب سوى تقديرات أصل وليس النتيجة الفعلية للتقاطع. بينما يمكن استخدام مرشحات بلوم لهذا الغرض، تم استكشاف العديد من البدائل. الأكثر واعدة بينها هو التباديل المستقلة للتجزئة الحساسة (**MIPs - Min-Wise Independent Permutations**).
يتطلب ذلك قوائم معرفات المستندات الرقمية كقيم إدخال. أولاً، تطبق هذه الطريقة دالات التجزئة الخطية **k**، مع مكون عشوائي، على القيم التي تعطي كل قائمة جديدة من القيم. ثانياً، يتم فرز جميع قوائم **k** الناتجة، حيث يتم الحصول على قوائم **k** المسموح بها، ويتم أخذ القيمة الدنيا لكل من هذه القوائم وإضافتها إلى قائمة جديدة: متجه **MIP** بالحجم **k**. البصيرة الأساسية هي أن كل عنصر لديه نفس الاحتمال في أن يصبح الحد الأدنى للعنصر تحت التقليب العشوائي. تقوم الطريقة بتقدير التقاطع بين متجهين **MIP** من خلال أخذ الحد الأقصى لكل موضع في المتجهين. يشكل عدد القيم المميزة في المتجه الناتج مقسومًا على حجم ذلك المتجه تقديراً للتداخل بينهما. الميزة هي أنه حتى لو كانت متجهات **MIP** المدخلة بطول غير متساوٍ، فلا يزال من الممكن استخدام المواضع القليلة الأولى فقط من المتجهات للحصول على تقريب أقل دقة. أن **MIPs** أكثر دقة بكثير من مرشحات بلوم لهذا النوع من التقدير١٩.

٢ / ٢ تقليل طول قوائم النشر باستخدام كلمات مفتاحية شديدة التمييز

تتمثل إحدى الطرق البديلة لتقليل تكاليف نشر تقاطع القوائم لفهرس عمومي موزع في جعل القوائم نفسها أقصر. لتحقيق ذلك، بدلاً من بناء فهرس على مصطلحات مفردة ، يمكن للمرء بناء واحد على استعلامات متعددة الكلمات المفتاحية بالكامل. هذه هي الفكرة وراء الكلمات المفتاحية شديدة التمييز. لم يعد يتم نشر جميع المصطلحات في فهرس موزع عمومي، ولكن بدلاً من ذلك يتم إنشاء استعلامات متعددة الأجل من محتوى الملف الذي يميز تلك الوثيقة جيدًا عن الآخرين في المجموعة. النتيجة: المزيد

من المنشورات في الفهرس، لكن قوائم نشر أقصر. يوفر هذا حلاً لأحد العيوب الرئيسية لاستخدام جداول التجزئة الموزعة: تقاطع قوائم النشر الكبيرة[٢٠].

٢ / ٣ الحد من عدد النتائج المعالجة باستخدام مداخل k الأعلى

يمكن أن تؤدي معالجة مجموعة فرعية فقط من العناصر أثناء عملية البحث إلى تحقيق فوائد في الأداء: تقليل معالجة البيانات وتقليل زمن الاستجابة. يمكن استخدام خوارزميات متنوعة، ستتم مناقشتها أدناه، لاسترداد العناصر العليا لاستعلام معين دون الحاجة إلى حساب الدرجات لكل العناصر. يعد استرجاع العناصر الرئيسية منطقيًا حيث ثبت أن مستخدمي محركات البحث على الويب يفضلون الجودة على الكمية فيما يتعلق بنتائج البحث: مزيد من الدقة وأقل استدعاء. تم تطبيق مداخل k الأعلى على مختلف الهياكل وفي مراحل مختلفة في استرجاع المعلومات من نظير إلى نظير وذلك على النحو التالي:

٢ / ٣ / ١ أعلى نتائج k المطلوبة

هناك طريقة بسيطة لتحسين النظام تتمثل في طلب أفضل النتائج فقط. تطبق المقاربات التي تستخدم الفهارس المحلية دائمًا شكلًا متغيرًا من النتائج المحدودة يطلب ضمنيًا عن طريق ربط عدد القفزات التي تم إجراؤها عند الفيضان أو عن طريق السير العشوائي الذي ينتهي. ومع ذلك، يمكن أيضًا تعيين هذا الرقم بشكل صريح على ثابت من قِبل الطالب كما هو الحال مع الفهرس المكرر الشامل الذي يستخدمه سوينكا-اكونا وآخرون (٢٠٠٣). حيث يحصلون أولاً على قائمة بنتائج البحث k ويواصلون الاتصال بالعقد طالما ظلت فرصة المساهمة في هذه الفئة مرتفعة. النتائج العليا تستقر بعد بضع جولات٢١.

٢ / ٣ / ٢ معالجة استعلام k الأعلى

هذا المنهج له جذوره في ابحاث البحث في قواعد البيانات. توجد عدة أشكال من هذا المدخل، جميعها لها نفس الفكرة الأساسية: يمكننا تحديد ملفات k العليا التي يتم تقديمها لعدة قوائم للمدخلات دون الاضطرار إلى فحص هذه القوائم تمامًا دون التأثير سلبًا على الأداء. يستخدم هذا غالبًا في الحالات التي يتم فيها استخدام فهرس عمومي موزع ويتم تقاطع قوائم النشر. خوارزمية العتبة هي الأكثر شعبية بين هذه المداخل. تحتفظ هذه الخوارزمية بنيتين للبيانات: قائمة انتظار مع أقرانهم للاتصال بهم للحصول على نتائج البحث وقائمة تحتوي على أفضل النتائج الحالية. تتم معالجة أقرانهم في قائمة الانتظار واحدًا تلو الآخر، ويعود كل منهم بمجموعة محدودة من نتائج بحث k للنموذج (الملف، الدرجة) مرتبة حسب العلامة بترتيب تنازلي. بالنسبة إلى فهرس عمومي موزع، فإن هذه هي أهم العناصر في قائمة النشر لفترة معينة. تتبع الخوارزمية مجموعتين لكل وثيقة فريدة: الأسوأ والأفضل. أسوأ نتيجة هي مجموع درجات مستند d الموجود في جميع قوائم النتائج التي ظهر فيها d. أفضل

نتيجة هي أسوأ نتيجة بالإضافة إلى أقل درجة (في بعض الملفات الأخرى) التي تمت مواجهتها في قوائم النتائج التي لم يظهر فيها d. نظرًا لأن جميع قوائم النتائج يتم اقتطاعها، فإن هذه النتيجة الأخيرة تشكل الحد الأعلى لأفضل درجة ممكنة والتي يمكن تحقيقها للوثيقة d. يتكون أعلى k الحالي من أعلى وثائق التسجيل التي شوهدت حتى الآن بناءً على أسوأ نتيجة لها. إذا كانت أفضل درجة في المستند أقل من العتبة، والتي تعتبر أسوأ درجة في المستند في الموضع k في نتائج أعلى الصفحة الحالية، فلا يلزم النظر إليها في أعلى الصفحة. وبالتالي فإن الخوارزمية تقوم على أساس التقاطع النهائي على نتائج k الأعلى فقط من كل نظير، مما ينتج عنه أداء مكافئ تقاطعًا "متسلسلًا" للقوائم بأكملها. هذا يحفظ كلاً من عرض النطاق الترددي والتكاليف الحسابية دون التأثير سلبًا على جودة النتائج. العيب هو أن البحث عن نتائج الوثيقة يتطلب وصولاً عشوائيًا إلى قوائم النتائج. قام زانج وسيول (٢٠٠٥)٢٢ في وقت لاحق بالتحقيق في مزيج من معالجة استعلام k الأعلى مع العديد من تقنيات التحسين. يخلصون إلى استنتاج مهم مفاده أن عمليات التحسين المختلفة قد تكون مناسبة لاستعلامات ذات أطوال مختلفة. وأظهر بالك وآخرون (٢٠٠٥)٢٣ أن معالجة استعلام k الأعلى يمكن أن تكون فعالة أيضًا في شبكات نظير إلى نظير مع مؤشرات محلية مجمعة.

٢ / ٣ / ٣ تخزين نتائج k الأعلى

هناك خطوة أخرى تتمثل في تخزين نتائج k الأعلى فقط للاستعلام أو المصطلح في الفهرس. يأخذ سكوبلاستين وابرير (٢٠٠٦)٢٤ هذا النهج كوسيلة لزيادة تقليل استهلاك حركة المرور. يرتبط هذا بنهج تانج ووركاداس (٢٠٠٤)٢٥ الذي يخزن المنشورات فقط من أجل المصطلحات العليا في المستند. يذكرون أنه على الرغم من أن فهرسة هذه المصطلحات العليا فقط قد تؤدي إلى تدهور جودة نتائج البحث، إلا أنه من غير المهم أن هذه الوثائق لن تحتل مرتبة عالية بالنسبة لاستعلامات المصطلحات الأخرى غير العليا التي تحتويها على أي حال.

٢ / ٤ تقليل عدد الأقران المشاركين في بحث الفهرسة بواسطة الفهرس الرئيسي المتكرر

تعتبر عمليات البحث عن خرائط لاستعلامات نظرائهم باهظة التكلفة عندما تتضمن الاتصال بأقرانهم الآخرين بغض النظر عن البنية المستخدمة. ماذا لو تمكن النظير من القيام بجميع عمليات البحث محليًا؟ يستكشف مؤلفو نظام PlanetP هذا المنهج الجديد٢٦. يتم فهرستها بالكامل إلى أقرانهم، وعناوين IP الخاصة بهم، وحالة شبكتهم الحالية، ومرشحات بلوم الخاصة بهم للمصطلحات. تنتشر هذه المعلومات عبر الشبكة باستخدام ثرثرة. إذا تغير شيء ما على مستوى النظراء، فمن المحتمل أن يغير طريقة

عمله. كل نظير يستقبل الشائعات وينشرها بنفس الطريقة. هناك احتمال أن يفوت أحد الأقران القيل والقال. في حين أن هذه طريقة مثيرة للاهتمام لنشر معلومات الفهرسة، إلا أنها بطيئة أيضًا بالاخذ في الاعتبار حقيقة أنها شبكة من تكنولوجيا المعلومات. وقد تم اعتماد هذا المنهج من قبل زينالبور - يازتي وآخرون (٢٠٠٤)٢٧.

على الرغم من أننا نفضل تسمية هذا المنهج كمؤشر عمومي، إلا أنه يمكن النظر إليه أيضًا على أنه شكل قوي جدًا من التكرار في الشبكة. لاحظ أن هذا المنهج يختلف عن جدول تجزئة موزع على قفزة واحدة، لأنه لا يحتوي على مفتاح تجزئة ولا يحتوي على مساحة مفاتيح موزعة. وبالتالي، لا يتم تحديد طبولوجيا الشبكة من خلال مفتاح المساحة keyspace.

٢ / ٥ تقليل تحميل المعالجة عن طريق التخزين المؤقت لنتائج البحث

من غير المنطقي إعادة بناء مجموعة نتائج البحث لنفس الاستعلام مرارًا وتكرارًا إذا لم يتغير هذا الاستعلام. يمكن زيادة الأداء بشكل كبير عن طريق تخزين نتائج البحث مؤقتًا. استخدم سكوبلاستين وابرير (٢٠٠٦)٢٨ جدول التجزئة الموزعة لتتبع الموضوع. في البداية، يكون هذا الجدول فارغًا، ويتم بث كل استعلام (متعدد الأجل) أولاً من خلال شبكة نظير إلى نظير بأكملها وذلك باستخدام بث مباشر لشبكة من الأقران. بعد هذه الخطوة في البحث عن نتائج الاستعلامات يسمح هذا باستهلاك البحث الفرعي: إرجاع نتائج البحث للمجموعات الفرعية للاستعلام في حالة عدم وجود تطابق كامل. يبني المؤلفون محتوى الفهرس على الاستعلامات التي يتم طرحها داخل الشبكة، وهي طريقة يطلقون عليها الفهرسة القائمة على الاستعلام. وينتج عن ذلك انخفاض حركة مرور الشبكة للاستعلامات الشائعة مع الحفاظ على نتيجة عمومية تتكيف في الوقت الفعلي مع استعلامات معينة.

٢ / ٦ تقليل عدد الأقران المشاركين في معالجة الاستعلام عن طريق التجميع

عند استخدام الفهارس المحلية، فإن إبقاء الأقران مع محتوى مشابه قريبًا من بعضهم البعض يمكن أن يجعل معالجة الاستعلام أكثر فاعلية. بدلاً من إرسال استعلام إلى جميع الأقران، يمكن إرساله إلى مجموعة من الأقران التي تغطي موضوع الاستعلام. هذا يقلل من العدد الإجمالي للنظراء الذين يجب الاتصال بهم لاستعلام معين. لسوء الحظ، لا تحدث المجموعات القائمة على المحتوى بشكل طبيعي في شبكات النظير إلى نظير. لذا حاول باوا وآخرون (٢٠٠٣)[٢٩] تنظيم شبكات نظير إلى نظير باستخدام تجزئة الموضوع. يقومون بترتيب نظرائهم في الشبكة بطريقة تحتاج فقط إلى استشارة مجموعة فرعية صغيرة من الأقران، تحتوي على مطابقة الملفات ذات الصلة، لاستعلام معين. يتم تنفيذ أقران المجموعات على أساس متجهات الملفات أو متجهات المجموعة الكاملة. ثم يستخدمون عملية من خطوتين لتوجيه الاستعلامات بناءً على الموضوع الذي يطابقونه. هم أول مجموعة من الأقران المسؤولة عن

موضوع محدد وإعادة توجيه الاستعلام هناك. بعد هذا، يتم إغراق الاستعلام داخل الكتلة الموضعية للحصول على التطابقات. وخلصوا إلى أن بنيتهم توفر أداءً جيدًا، من حيث جودة الاسترجاع ومن حيث وقت الاستجابة وعرض النطاق الترددي. لسوء الحظ، يحتاج نظامهم إلى فهرس مركزي للعثور في البداية على مجموعة مواضيع جيدة لتوجيه الاستعلام.بينما عرض كافيبات وآخرون (٢٠٠٦)[٣٠] كيفية القيام بالتجميع دون مثل هذا الفهرس المركزي.

قام كلامبنوس وخوزيه (٢٠٠٧)٣١ بتقييم البنية القائمة على نظام المجموعة لاسترجاع المعلومات من نظير إلى نظير على نطاق واسع مع التركيز على التجميع الفردي مع كل من العدد المتغير والثابت من المجموعات. ووجدوا أن الحجم الصغير في الغالب لملفات الويب يجعل من الصعب ربطهم بالوثائق الأخرى مما يؤدي إلى ضعف المجموعات. تفشل آليات التجميع في اكتشاف بنية توزيع الملف الأساسي الذي يؤدي إلى الموقف حيث لا يمكن الاتصال بالمصادر ذات الصلة الكافية لتوجيه الاستعلام. هذا يرجع إلى فقدان المعلومات الكامنة في إنشاء النقط الوسطى العنقودية. لذا يقترحون حلين. أولاً، نسخ المستندات جزءًا من مجموعات شعبية على أقران متعددين، مما يؤدي إلى تحسن كبير في الفعالية. على الرغم من أن هذا لا يحل المشكلة بالنسبة للمواضيع التي لا تحظى بشعبية، إلا أنه قد يعمل بشكل جيد بما فيه الكفاية لمعظم المستخدمين. ثانياً، بافتراض وجود آلية تغذية مرتدة ذات صلة واستخدام ذلك لتغيير الأقطار الوسطى لمجموعات الموضوعات العمومية. ثم يتم تحديد وزن كل مصطلح في مجموعة حسب مدى صلة تلك المجموعة بالاستعلام بناءً على الملاحظات. إنها تُظهر فائدة كل من ردود الفعل المتعلقة بالنسخ المتماثل والأهمية والتي تؤدي إلى توجيه أفضل للاستعلام ودقة أعلى، مع التركيز بشكل أكبر على التعليقات ذات الصلة كتطور واعد وطبيعي لتقنيات استرجاع المعلومات من نظير إلى نظير.

تعمل المجموعات القائمة على الاهتمامات إما عن طريق تقصير أطوال المسار بين الأقران ذوي الاهتمامات المتشابهة، أي: الأقران الذين يطرحون استفسارات متشابهة، أو عن طريق تقريب الأقران ذوي الاهتمامات الخاصة من أقرانهم بمحتوى مطابق. على الرغم من أن هذا ليس هو نفسه تمامًا، فإن كلاهما يهدف إلى تقليل عدد القفزات اللازمة للحصول على محتوى ذي صلة. في الحالة الأولى عن طريق الاستفادة من المعلومات المخزنة مؤقتًا الموجودة في الأقران ذوي الاهتمامات المتشابهة، في حين أن الحالة الثانية تقرب المرء من أصل المعلومات: توفير الأقران التي تحتوي على محتوى أصلي. ويمكن أيضًا الجمع بين نهجين المجموعات، حسب المحتوى والفائدة٣٢.

٢ / ٧ الحد من وقت الاستجابة وتحسين الاستدعاء باستخدام المسار العشوائي

يتم عادةً البحث عن نظم نظير إلى نظير مع مؤشرات محلية باستخدام فيضان الاستعلام. هذا المنهج شامل من الناحية النظرية، ولكن بسبب قابليته للتتبع، يتم تطبيقه بطريقة غير شاملة عن طريق ربط عدد القفزات. اقترح لاف وآخرون (٢٠٠٢)٣٣ بديلاً لذلك باستخدام مسارات عشوائية. بدلاً من البحث بطريقة أولية متسعة: إعادة توجيه الاستعلامات إلى جميع الأقران، فإننا نبحث في العمق أولاً: عن طريق إعادة توجيه الاستعلام إلى قرين واحد فقط. تتبع عملية السير هذه من نظير الاستعلام، والخطوات العشوائية للأمام عبر الشبكة. الأقران الذين لديهم النتائج ذات الصلة إرسال هذه مرة أخرى مباشرة إلى النظير الأصلي. يقوم الأقران المشاركون في المسيرة أحيانًا بالتحقق من الرضا لدى النظير الأصلي فيما يتعلق بعدد النتائج التي تم الحصول عليها حتى الآن وإنهاء المسيرة بناءً على ذلك. رغم أن أن هذا المنهج بطيء، ولكن يمكن بدء مسارات متعددة بشكل متوازٍ لتقليل وقت الاستجابة. شبيه بالمسار العشوائي اقترح كالوجيراكي وآخرون (٢٠٠٢)٣٤ إعادة توجيه رسائل الاستعلام إلى جزء تم اختياره عشوائيًا من الأقران المجاورة.

٢ / ٨ الحد من وقت الاستجابة وتحسين الاستدعاء باستخدام المسار المباشر

قام آداميك وآخرون (٢٠٠١)٣٥ بتوجيه رسائل الاستعلام عبر العقد ذات الدرجة العالية: تلك التي تتمتع باتصال عالي، وأتضح أن هذا يقلل من وقت البحث ويزيد من تغلغل الشبكة. بينما قام يانج وجارسيا-مولينا (٢٠٠٢)٣٦ بإعادة توجيه رسائل الاستعلام إلى الأقرانهم الذين سبق لهم أن عرضوا معظم نتائج الاستعلام. في سياق مشابه قدم روسبولوس تسوماكوس (٢٠٠٣)٣٧ البحث الاحتمالي التكيفي حيث يحتفظ كل من الأقران بجدول توجيه احتمالي لكل استعلام نشأ منه أو انتقل من خلاله. يقوم النظير الأولي الذي يرسل استعلامًا ببثه إلى جميع أقرانه، ولكن من هناك في رسالة الاستعلام يتم توجيهه فقط إلى النظير الذي لديه أعلى احتمال للحصول على النتائج بناءً على التعليقات السابقة. وبناءً على ذلك يقترحا آلية يقوم فيها الأقران ببناء ملفات تعريف للنظراء المجاورين بناءً على أحدث الاستفسارات التي استجابوا لها، يسجل النظير الاستعلامات الواردة مقابل ملفات تعريف اقرانه مرتبة حسب النوعية، بناءً على تشابه جيب التمام الخاص بهم، وكميًا: عدد النتائج التي تم إرجاعها مسبقًا. هذا يتفوق على الفيضان الأساسي، وإعادة التوجيه العشوائي، والتوجيه الكمي الخالص الموجه.

بينما قام تشونغ وآخرون (٢٠٠٣)٣٨ بمقاربة اقتصادية للاستعلام عن التوجيه في شبكات الهاتف المحمول باستخدام نظام ائتماني بسيط ومضاد للغش يعتمد على سلطة ائتمان مركزية تفرض رسومًا على الأقران لإرسال الرسائل. يمكن الحصول على الائتمان بأموال حقيقية أو عن طريق إعادة توجيه الرسائل إلى الأقران الآخرين.

واعتبر لى وآخرون (٢٠٠٩)٣٩ هذا العمل مصدر إلهام وقاموا بتطبيقه على شبكات نظير إلى نظير ذات المؤشرات المحلية. عندما يصدر أحد الأقران استعلامًا، فإنه يربح مقابل النتائج. على وجه التحديد، يتم وعد هذه الأقران المجاورة المجاورة كدفعة عند إرجاع نتائج البحث ذات الصلة عن طريقهم. يمكن للأقران اختيار النظراء الآخرين الذين يقومون بإعادة توجيه استعلام والقيام بذلك مقابل جزء من القسط المقدم لهم. أخيرًا، عندما تكتشف عملية توجيه الاستعلام نظيرًا له نتائج بحث ذات صلة، فإنه يعيدها إلى مسار النظير الذي بدأ الاستعلام. على طول الطريق يحصل كل نظير على المكافأة الموعودة. يمكن استخدام عملة المكافأة هذه لإصدار استعلامات جديدة لكل نظير، وبالتالي تشجع على المشاركة في التوجيه. أظهر المؤلفون أن مقاربتهم تستخدم بشكل أفضل قدرة شبكة نظير إلى نظير من استخدام فيضان الاستعلام والمسار العشوائي.

المبحث الثالث: نظم استرجاع المعلومات من نظير إلى نظير

تم تطوير العديد من نظم استرجاع المعلومات من نظير إلى نظير لتطبيقات مختلفة. غالبًا ما تستعير هذه النظم عناصر من شبكات مشاركة البيانات واسترجاع المعلومات الموحدة بمستويات مختلفة من النجاح. تركز معظم نظم البحث على مجال شبكات الحوسبة أو المكتبات الرقمية أو الويب.

وعلى الرغم من وجود العديد من نظم البحث، فنحن نقصر أنفسنا على مجموعة فرعية منها في هذا القسم. نناقش ال النظم برزت بسبب طبيعتها الرائدة أو عن طريق استخدامها لمزيج مثير للاهتمام من التقنيات.

Sixearch ١ / ٣

كان مشروع Infrasearch أحد أوائل نظم استرجاع المعلومات من نظير إلى نظير، والذي أصبح فيما بعد JXTASearch الذي يعد أيضًا الأساس لـ Sixearch. يتكون Sixearch من عدة مكونات: نظام توثيق الملفات، ومحرك استرجاع، و JXTA للاتصال بشبكة نظير إلى نظير، ونظام لتعلم المحتوى. يستخدمون بنية تستند إلى XML ويفترضون أن الاستعلام يتكون من مجموعة قابلة للتخصيص منظمة من الحقول. يمكن أن تحتوي مجموعة الكتب على سبيل المثال على الحقول: العنوان، المؤلف، إلخ. لا يتم توجيه هذا المنهج جيدًا نحو استرجاع النص الكامل لأنه يعتمد على بنية الاستعلامات بدلاً من بنية المحتوى المشترك. ينشر الأقران الموردين، لكل حقل استعلام، مجموعة من الكلمات الرئيسية التي يعتقدون أنها يمكن أن توفر نتائج ذات صلة: وصف المورد الخاص بهم. يطرح أقران المستهلكين استعلامات منظمة يتم توجيهها بعد ذلك إلى أقران الموردين المناسبين باستخدام المحاور. يعتمد توجيه الاستعلام على ملفات تعريف المحتوى الخاصة بالأقران المجاورة التي يتم تحسينها

باستمرار باستخدام تعلم التعزيز القائم على التفاعلات السابقة. يستخدم دمج نتائج البحث التي تم إرجاعها بواسطة عدة أقران خوارزمية فهرسة بسيطة. يريد مصممي النظام تحسين نظامهم من خلال التركيز على التعلم السياقي والتعاون الاجتماعي. ويعتزمون توسيع نظامهم بنظام سمعة كمكون أمان لتمييز مرسلي البريد العشوائي عن أقرانهم الصادقين[٤٠].

٣ / ٢ ODISSEA

يتكون هذا النظام من مستويين. يتكون المستوى الأدنى من الأقران الذين يحتفظون بمؤشر رئيسي موزع. توجد دائمًا منشورات المصطلح في نظير واحد. يتكون المستوى العلوي من أقران التحديث الذين يقومون بإدراج أو تحديث الملفات في النظام، مثل خادم الويب، والاستعلام عن الأقران الذين يستخدمون الطبقة الدنيا للإجابة على الاستفسارات. الحداثة في مقاربتهم هي في تخصص الأقران وكذلك في استخدام لمفهوم وثيقة المصطلح العمومي الموزع. تجعل التخصصات الأقران المسؤولين عن تخزين وبناء واستعلام الفهرس في نظامهم هم في الواقع مفصولون. يشبه هذا المنهج المركزي تمامًا الذي تستخدمه عادة محركات البحث الحديثة حيث تقوم بعض الأجهزة فقط بتخزين الملفات في الفهرس، بينما يزحف البعض للحفاظ على الفهرس جديدًا، بينما يقوم الآخرون (الخارجيون) بالاستعلام فقط. على النقيض من تلك النظم، توفر ODISSEA بنية تحتية للفهرسة والبحث مفتوحة حيث يمكن لكل جهاز يشترك في البروتوكول أن يشارك كنظير[٤١].

عند التعامل مع الاستعلامات متعددة الأجل، تتم تقاطعات قائمة الترحيل بترتيب تصاعدي لحجم قائمة نشر المصطلحات: من الصغيرة إلى الكبيرة، لأن هذا يقلل بدرجة كبيرة من كمية البيانات التي يجب نقلها. علاوة على ذلك، يطبق النظام معالجة استعلام k الأعلى لتقليل استخدام عرض النطاق الترددي. يقترح مصممي النظام تحسين أساليب تنفيذ الاستعلام. علاوة على ذلك، فقد ذكروا أن استرجاع المعلومات على نطاق الويب يعد تطبيقًا أكثر صعوبة من مشاركة الملفات.

٣ / ٣ MINERVA

يفترض هذا النظام أن كل نظير يقوم بتنفيذ عمليات الزحف الخاصة به ويقوم بإنشاء فهرس محلي. يبحث النظير أولاً في الفهرس المحلي الخاص به للعثور على نتائج البحث ذات الصلة. إذا كانت هذه النتائج غير مرضية، فيمكن للنظير الرجوع إلى فهرس رئيسي موزع يحتوي على قائمة من الأقران الذين لديهم ملفات ذات صلة لكل فهرس في الشبكة: فهرس من خطوتين. يحتوي هذا المؤشر العمومي أيضًا على إحصائيات تتعلق بالمؤشرات المحلية التي يحتفظ بها كل نظير. يوضح مصممو النظام أن تقدير التداخل بين نتائج البحث بشكل صحيح يمكن أن يقلل من عدد الأقران الذين يحتاجون إلى الاتصال بهم لاسترجاع الاستدعاء الكامل بأكثر من ستين بالمائة.

ومع ذلك، تظل عمليات البحث في جدول التجزئة الموزعة باهظة الثمن. يقترح المصممون استخدام الارتباطات بين المصطلحات الفردية في الاستعلام لتقليل عدد عمليات البحث: يقوم النظير الذي يعالج المصطلح الأول في الاستعلام أيضًا، بشكل تكيفي، بتخزين ما أقرانه للاتصال بشروط الاستعلام المتبقية. هذا يقلل بشكل كبير من عدد الأقران المشاركين، مع الحفاظ على نفس المستوى من الاستدعاء[٤٢].

ومع ذلك، يمكن أن تتسبب المصطلحات الشائعة في حدوث اختلالات حادة في التحميل إذا كان النظير الواحد يتحمل مسؤولية تخزين جميع المنشورات لمدة واحدة. اقترح المصممون إنشاء فهارس مؤلفة من وثيقة خطوة واحدة في MINERVA للمصطلحات الشائعة لتقليل أوقات الاستجابة. نظرًا لأن قوائم النشر عادةً ما يتم مسحها ضوئيًا بالتتابع، من الأفضل إلى الأسوأ في تصنيف النقاط لمدة معينة، فإنها تستخدم أمرًا يحتفظ لتخزين المنشورات لمدة مصطلح يتم فرزها عن طريق تنازلي على نقاط متعددة. يطبق المؤلفون معالجة استعلام k الأعلى لتقليل الحمل. يمكن تحسين ذلك بشكل أكبر عن طريق تطبيق التخزين المؤقت لنتائج البحث: تخزين نتائج البحث المخزنة مؤقتًا لكل استعلام كامل على الأقران الذين يقومون بتخزين المنشورات لأحد مصطلحات الاستعلام. تحتوي هذه النتائج على معلومات التعريف التي تساعد في تحديد ما إذا كانت لا تزال حديثة بما فيه الكفاية ومن يجب الاتصال به للحصول على نتائج محدثة. يوضح مصممي النظام أن تدوير العنصر الأقل استخدامًا هو أفضل استراتيجية لإدارة التخزين المؤقت لذاكرة تخزين مؤقت محددة. يقوم مصممو النظام بتجربة كل من التخزين المؤقت الدقيق: مطابقة استعلام متعدد المدة تمامًا، وذاكرة التخزين المؤقت التقريبية: مطابقة مجموعات فرعية مصطلح استعلام متعدد المصطلحات. وجدوا أن كلا النهجين يوفران موارد شبكة قيمة دون المساس بجودة النتائج[٤٣].

وبالنظر في حداثة الملفات الإضافية الناتجة بالإضافة إلى الجودة الخالصة باستخدام خوارزمية اختيار مجموعة استرجاع المعلومات المعدلة. يتم التركيز على تحسين توجيه الاستعلام. علاوة على ذلك، يقومون بتجربة مرشحات بلوم والتباديل المستقل Min-Wise، مما يدل على أن الأخير مناسب بشكل أفضل للحصول على تقديرات حجم مجموعة النتائج.

٣ / ٤ ALVIS

ألفيس هو منهج المؤشر العمومي الموزع، مع العديد من الابتكارات. أثناء استرجاع النتيجة النهائية، يتم الاتصال بكل نظير قام بإنشاء إدخال فهرس ويطلب منه إعادة احتساب درجة الوثيقة بناءً على إحصاءات عمومية ومحلية، وبالتالي توليد درجات قابلة للمقارنة بشكل عام. بدلاً من تخزين منشورات للمصطلحات الفردية، يستخدم

المصممون كلمات مفتاحية شديدة التمييز. يقدم هذا مشكلة الحاجة إلى تخزين العديد من الكلمات المفتاحية أكثر من الفهرسة التقليدية. للتخفيف من ذلك، قاموا بجمع أسلوبهم مع الفهرسة التي تعتمد على الاستعلام وتخزين الكلمات المفتاحية الشائعة فقط في الفهرس وتطبيق تخزين أفضل النتائج. في حين أن هذا لا يناسب على استفسارات أقل شعبية، أظهر بالفعل أن سجلات الاستعلام يمكن استخدامها بفعالية لضبط الكلمات المفتاحية غير الملائمة من الفهرس دون فقد الكثير من جودة الأداء[٤٤].

PHIRST ٥ / ٣

تؤدي الاختلافات بين المؤشرات العمومية والمحلية إلى ظهور خيارات متبادلة. علينا أن نختار بين البحث السريع، ولكن المكلف وغير المرن، وبين البحث الدقيق أو البحث البطيء، ولكن الرخيص والمرن، التقريبي. يقدم هذا النظام أسلوب للبحث عن النص الكامل للنظير يجمع بين المؤشرات العمومية والمحلية لأنواع مختلفة من المصطلحات. يحتفظون فقط بمصطلحات التردد المنخفض في جدول التجزئة، بينما يقومون بتقدير عدد المصطلحات الشائعة من خلال فيضان الاستعلام المحدود. من المحتمل أن تحتوي الملفات التي تمت إضافتها حديثًا على مصطلحات أكثر تكرارا معروفة وأقل مصطلحات جديدة منخفضة التكرار. وبسبب هذا التأثير، يزعم مصممي النظام أن مقاربتهم تؤدي إلى فهرس أصغر نسبيًا مع زيادة عدد الملفات والأقران المفهرسة مقارنةً بفهرس كامل محفوظ في جدول تجزئة موزع. ظهر بالفعل أن هذا النهج المختلط يحسن أوقات الاستدعاء والاستجابة ويتحمل مقدار أقل من عرض النطاق الترددي للبحث في مشاركة الملفات[٤٥].

Klampanos2004 ٦ / ٣

يحاول هذا النظام تطبيق أساليب استرجاع المعلومات القياسية في بيئة نظير إلى نظير. وهي تجمع بين المؤشرات المحلية المجمعة وتجميع المحتوى. ويفترضون أن كل نظير يفهرس ملفاته الخاصة ويجد مجموعات المحتوى في مجموعته الخاصة. على مستوى الشبكة، ينضم كل نظير إلى مجموعة واحدة أو أكثر من المجموعات التي تدرك المحتوى بناءً على المجموعات المحلية. المجموعات التي تدرك المحتوى هي، على الأرجح، مجموعات من الأقران. يقوم كل نظير فائق بتخزين واصفات كل مجموعة من هذه المجموعات في الشبكة ويمكن أن يسجلها ضدهم في حالة وجود استعلام. يتم استخدام نسخة مبسطة لنظرية الأدلة (Dempster-Shafer)، وهي طريقة لدمج الأدلة من مصادر متعددة في اعتقاد مشترك واحد، وفي دمج النتائج التي قدمها نظراء متعددين. يبدو أن هذا جيد الأداء، أثناء الاتصال بعدد قليل من الأقران: عادةً واحد أو اثنين، وأحيانًا ثلاثة ونادراً ما يكون ستة[٤٦].

٣ / ٧ NeuroGrid

ظهر نظام بحث لامركزي تكيفي يسمى NeuroGrid يستخدم فهرسة هجينة: في البداية يكون لدى جميع الأقران فهرس ملفات محلي خاص بهم ، لكن عندما ينضمون إلى الشبكة، يقومون بإنشاء فهرس نظير من الأقران المجاورين. يشبه هذا عن كثب المؤشرات المحلية المجمعة، ولكن مع كل نظير يعمل كنظير فائق. في البداية شبكة NeuroGrid هي عبارة عن شبكة من رسائل الفيضانات البسيطة. الحداثة في المنهج هي في التوجيه التكيفي للاستعلامات. يتم تسجيل ردود المستخدمين على نتائج البحث، وعدم وجود ردود فعل إيجابية أو ردود فعل سلبية واضحة. عندما يتعين على NeuroGrid تحديد مجموعة فرعية من الأقران لإعادة توجيه استعلام إليها، فإنه يحاول زيادة فرصة تلقي تعليقات إيجابية للنتائج التي تم إرجاعها بناءً على هذه التجارب السابقة. في حالة وجود ردود فعل إيجابية، ينشئ نظير الاستعلام رابطًا مباشرًا للنظير المستجيب في شبكة التراكب. يزيد هذا النوع من التجميع تدريجياً من الاتصال ويجعل جميع الأقران أكثر دراية فيما يتعلق بمحتوى جيرانهم. يقلل هذا المنهج أيضًا من طول المسار الذي تحتاجه الاستعلامات في الانتقال عبر الوقت. النظام يفضل الأقران الموثوق بهم: أولئك الذين يستجيبون للاستعلامات ويقدمون نتائج موضوعية تهم المستخدم. أقرانهم ذوو الصلة الجيدة هم أكثر تأثيراً على عملية التعلم الإحصائي[٤٧].

٣ / ٨ Galanis2003

يقترح هذا النظام تنظيم جميع مصادر البيانات على الإنترنت في شبكة كبيرة من نظير إلى نظير حيث يتم الرد على الاستفسارات من قبل المواقع ذات الصلة. يفترض مصممو النظام أن كل نظير هو محرك بحث XML يحافظ على فهرس محلي. عندما ينضم نظير إلى الشبكة، يرسل إلى أقرانه الآخرين ملخصًا لبياناته: مجموعة صغيرة من العلامات المختارة وممثل الكلمات الرئيسية لمحتواها. إن الرابط يولد بالتالي موجة من الرسائل، مما يجعل منهجهم موجهًا نحو الشبكات ذات الاضطرابات المنخفضة جدًا. بدلاً من ذلك يمكن أيضًا تخزين هذه المعلومات عند إرسال الاستعلامات. يكتسب الأقران أيضًا، في البداية من أقرانهم المجاورين، ملخصات محتوى لأقرانهم الآخرين في النظام ويحافظون على فهرس النظراء الخاص بهم. اختبر مصممي النظام مع تكرار الملخصات لكل نظير وللنظير مجموعات فرعية من مختلف الأحجام. تشير نتائجهم إلى أن استخدام النسخ المتماثل لكل نظير يتفوق على استخدام المجموعات الفرعية، رغم أن استخدام مجموعات فرعية كبيرة يمكن أن يقترب من هذا الأداء. يقارنون مؤشرات التجميع الكاملة للنسخ المتماثل بنهج

المؤشرات المحلية الصارمة ويظهرون أن التجميع يزيد من إنتاجية الاستعلام بنسبة ٢٠٧١٪ ويوفر أوقات استجابة أسرع ٧٢ مرة[٤٨].

٣ / ٩ Triantallou2003

في ظل هذا النظام يتم التركيز على فرض توزيع الحمل العادل بين الأقران. من خلال القيام بتجميع المستندات في فئات دلالات وأقران المجموعات بناءً على فئات المستندات التي يقدمونها. يؤكد مصممو النظام على الحاجة إلى فرض بعض بنية النظام المنطقي لتحقيق أداء عالٍ في شبكة استرجاع معلومات نظير إلى نظير. يحتفظ الأقران بفهرس المستند، الذي يعين معرّفات المستندات للفئات، وفهرس الكتلة الذي يعين الفئات إلى معرّفات الكتلة، وفهرس النظير الذي يعين معرّفات الكتلة للأقران. يتم أولاً تعيين المصطلحات في استعلام إلى فئات، ثم إلى مجموعات وأخيراً إلى نظير عشوائي داخل المجموعات ذات الصلة. يحاول هذا النظير العشوائي تلبية الاستعلام من خلال النتائج المحلية الخاصة به، ولكن في حالة توفر عدد قليل للغاية، فإنه يعيد توجيه الاستعلام إلى العقد المجاورة في نفس المجموعة. هذا يتكرر حتى تكون هناك نتائج كافية. نظرًا لأن الاختيار عشوائي، فمن المحتمل أن يتم انتقاء كل نظير في المجموعة مما يحقق موازنة التحميل بين أقرانهم داخل المجموعة نفسها. يتم تعيين الأقران إلى مجموعات بناءً على فئات المستندات التي يشاركونها. من أجل موازنة التحميل بين الكتل، يقدم المؤلفون مؤشر الإنصاف وخوارزمية فعالة لزيادة هذا الحد الأقصى المسموح به إلى MaxFair والذي يعوض أيضًا نظرائهم بقدرات معالجة مختلفة وتوزيع المحتوى والتخزين. تم تعيين أقوى نظير في كل مجموعة كقائد الذي يشارك في خوارزمية MaxFair. قد يتم إعادة تعيين الفئات ديناميكيًا إلى مجموعة مختلفة لتحسين الإنصاف استنادًا إلى تحميل كل مجموعة. يظهرون أن أسلوبهم قادر على الحفاظ على العدالة حتى عندما يتغير أقرانهم ومجموعات المستندات[٤٩].

المبحث الرابع: تحديات نظم استرجاع المعلومات في شبكات نظير إلى نظير

لقد رأينا بالفعل بعض التحديات التي تنطبق على استرجاع المعلومات في شبكات نظير إلى نظير بشكل عام خلال استعراض تقنيات ونظم استرجاع المعلومات المختلفة في المبحثين السابقين. في هذا القسم، نناقش مجموعة فرعية من هذه التحديات الأكثر أهمية لاسترجاع معلومات نظير إلى نظير والتي يجب اخذها في الحسبان عند تصميم اي اداة او نظام في المستقبل.

٤ / ١ وقت الإستجابة

في استرجاع معلومات نظير إلى نظير، يتم التحكم في زمن الوصول الذي يحدث عند تنفيذ عمليات البحث بعدد الأقران المشاركين في توجيه الاستعلامات ومعالجتها. لقد رأينا أن المؤشرات المحلية والعمومية مناسبة لأنواع مختلفة من الاستعلامات، وأن

هناك العديد من عمليات التحسين التي يمكن تطبيقها لتقليل تكلفة تخزين ونقل معلومات الفهرس. ومع ذلك، يظل التحدي المتمثل في الجمع الأمثل بين هذه التقنيات، وإيجاد تقنيات جديدة، للحفاظ على وقت الاستجابة ضمن حدود مقبولة. السبب في ذلك في المقام الأول هو أنه لا يوجد حل واحد جيد لجميع الحالات وأن هناك كمية متزايدة من المعلومات للفهرسة مما يؤدي إلى مشاكل وقت استجابة أكبر. بالنسبة إلى أي نظام بحث، من المهم تقديم نتائج البحث بسرعة دون المساس بالكثير على جودة النتائج. الأسباب الفنية للتأخير لا صلة لها بالمستخدمين. بعد إدخال استعلام، يجب أن تظهر النتائج في أكثر من ثانيتين. ومع ذلك، من الناحية المثالية ينظر إلى النتائج على أنها تظهر بشكل فوري للتنافس مع الحلول المركزية، مما يعني تأخيرًا قدره ٠,١ ثانية. أي شيء دون ذلك من غير المرجح أن يؤثر بشكل إيجابي على تجربة المستخدم. تركز معظم الحلول الحالية بحق على تقليل عدد القفزات و / أو استخدام التوازي للحد من وقت الاستجابة. يمثل توجيه الاستعلام الفعال تحديًا محددًا لاسترجاع معلومات نظير إلى نظير ويرتبط مباشرةً بوقت الاستجابة. يمكن أن يؤدي المزيد من البحث في الجوانب الزمنية للاستعلام إلى حلول أكثر ملائمة[٥٠].

٤ / ٢ الحداثة

إن إبقاء الفهرس جديدًا يمثل تحديًا لكل محرك بحث. حل هذه المشكلة إلى حد كبير هو مهمة مكون تتبع ارتباطات الويب لمحرك البحث. يجب أن يكون الفهرس ممثلاً لمواقع الويب المفهرسة، دون تكبد حمل كبير على تلك المواقع لاكتشاف التغييرات. تتغير بعض ملفات الويب سريعًا ونادراً ما تتغير بعضها، وليس كل تغيير يحدث مهمًا بما يكفي لضمان تحديث الفهرس. في مواقع الحالة المثالية، تشارك مواقع الويب بشكل تعاوني في شبكة نظير إلى نظير وتشير إلى تغييرات مهمة على الشبكة. هذا سيزيل الحاجة إلى الزحف. ومع ذلك، سيتعين على محركات بحث الويب من نظير إلى نظير التعامل في البداية مع الموقف الحالي. يبدو أن قيام الأقران بأداء عملية الزحف الخاصة بهم يمثل طريقة واقعية، ولكنه يقدم نفس المشكلات التي تحدث مع الزحف التقليدي على الويب. نظرًا لأن العديد من التحديثات يمكن أن تحدث بسبب تغيير الملفات، فمن المهم أن يكون لأي مؤشر يتم استخدامه الحد الأدنى من طفرة الحمل. يمكن استخدام استراتيجيات فهرسة منفصلة لملفات ويب سريعة وبطيئة التغيير. وهناك تحد آخر للحداثة هو استخدام التخزين المؤقت الموزع لقوائم المنشورات أو نتائج البحث. هذه الآليات تقلل من وقت الاستجابة، ولكنها تفعل ذلك على حساب الحداثة[٥١].

٤ / ٣ التقييم

على الرغم من أن التقييم عن طريق المحاكاة يمكن أن تحتوي على العديد من المتغيرات الايجابية لشبكة نظير إلى نظير، إلا أنه من أجل المقارنة الدقيقة، يجب استخدام نفس البيانات. هناك حاجة إلى مجموعة مشتركة، وطريقة لتوزيع هذه المجموعة على الأقران ومجموعات الاستعلام. كانت هناك محاولتان على الأقل لإنشاء مثل هذا المعيار، على الرغم من أنهم لم يروا بعد تبني واسع النطاق. أن تقييم نظم استرجاع المعلومات من نظير إلى نظير مهمة شاقة ومهملة. أنها تنشئ عددا من أسرة اختبار وثيقة مختلفة لتقييم هذه النظم. يذكرون أن تقييم هذه الشبكات صعب للغاية بسبب عدة أسباب. أولاً، يفترض أن تكون كبيرة جدًا مما يجعل المحاكاة أكثر عبئا. ثانياً، يتعرضون للمشاكل الناتجة عن الأقران التي تتعطل. لسوء الحظ، لم يتم التحقيق بشكل جيد في تأثير الاستخراج في تجارب استرجاع المعلومات من نظير إلى نظير، حيث يفترض معظمهم بيئة دائمة الاستخدام. ثالثًا، من غير المحتمل أن يتم وضع المستندات بشكل عشوائي على الأقران، بدلاً من ذلك يتأثر توزيعها بالموقع السابق واسترجاعها والتكرار. أخيرًا، محاكاة سلوك المستخدم معقدة. على سبيل المثال: محاكاة واقعية لكيف تتغير مجموعات وترددات الاستعلام مع مرور الوقت. يتم التحايل على هذا عادةً عن طريق عكس السلوك في توزيع المستند. ومع ذلك، فمن الطبيعي أن تعكس سيناريو التطبيق إلى الحد الذي يمكن أن تكون فيه النتائج قاطعة[٥٢].

توجد أنواع مختلفة من شبكات استرجاع معلومات نظير إلى نظير لها توزيعات مختلفة للوثائق. أولاً، مجال الويب الذي تتبع فيه توزيعات المستندات قانون الشبكة. ثانياً، الشبكات التي يتم التحكم فيها بشكل فضفاض مع توزيع موحد للملفات التي تفرض حملاً متساوياً على كل نظير. ثالثًا، المكتبات الرقمية حيث يتبع التوزيع أيضًا قانون الشبكة، على الرغم من أنه أقل تطرفًا من مجال الويب. بالإضافة إلى ذلك، تتميز المكتبات الرقمية بعدد أقل من الأقران التي تشترك كل منها في كمية أكبر بكثير من الملفات مقارنة بالأقران في الحالات الأخرى. يتم محاكاة النسخ المتماثل في جميع هذه المجالات من خلال استغلال الروابط بين المجال[٥٣].

المبحث الخامس: الدراسة التجريبية

فيما يلي نعمل على تطوير نظامًا مبنيًا على نظام لايم واير Gnutella. LimeWire هو برنامج مفتوح المصدر مكتوب بلغة جافا. يسمح للمستخدمين بمشاركة أي نوع من الملفات وتشغيله على نظم Windows و Macintosh و Linux و Sun وغيرها من منصات الحوسبة. على وجه التحديد، سنقوم بإضافة الوظائف التالية إلى عميل LimeWire:

١. صيانة الإحصاءات: يقوم كل عميل بجمع إحصائيات حول الملفات المشتركة.

٢. وظائف التصنيف لأهداف عمليات استرجاع المعلومات: يقوم كل عميل بتنفيذ عدد من وظائف الترتيب.
٣. مسجل البيانات: يمكن لكل عميل تسجيل البيانات الواردة والبيانات الصادرة مع النتائج.
٤. محلل البيانات: مكون مستقل يقوم بتحليل البيانات المسجلة بهدف تسهيل عمليات استرجاع المعلومات.

٥ / ١ وصف موجز لـ GNUTELLA

اخترنا أن نبني عملنا على بروتوكول Gnutella لأنه ليس فقط شائعًا ولكنه أيضًا خضع للبحث والتقييم والدراسة بشكل متعمق. ففي شبكة Gnutella، يبحث المستخدم عن ملف عن طريق إصدار استعلام كلمات رئيسية. ويتم إرسال هذا الاستعلام إلى جميع العملاء المتصلين به بنشاط ما (يكون عددهم عادةً صغيراً، في حدود ١٠). ثم يعيد العميل توجيه هذا الاستعلام إلى جميع جيرانه في شبكة Gnutella. تتكرر هذه العملية إلى أن تنتهي صلاحية Time-To-Live (TTL) للاستعلام أو تصل الحزمة إلى عميل يمثل عددًا محددًا مسبقًا من "القفزات" بعيدًا عن المرسل.

تحدد Gnutella شبكة غير منظمة وموزعة للغاية حيث تكون كل عقدة مستقلة تمامًا، وتتحكم بشكل مستقل في مستودعها المحلي للملفات المشتركة. تتم مقارنة الاستعلامات الواردة مع واصفات الملفات (نظرًا لأن هذه الملفات المشتركة هي ملفات ثنائية، فإنها تحتاج إلى واصفات خارجية). يتم تطبيق واصفات الملفات بشكل عام عبر أسماء الملفات. يتم إرجاع نتائج المطابقة إلى النظير الذي أصدر الاستعلام.

٥ / ٢ وصف أداة البحث

أهداف IR-P2P ذات شقين: أداة بحث وأداة إعادة بحث. فيما يلي وصف البنية العامة لنظامنا مع شرح وظيفة كل مكون.

٥ / ٢ / ١ بنية نظام IR-P2P

يمكن إيقاف تشغيل وظائف تسجيل IR-P2P إذا كانت وظائف LimeWire الأساسية مطلوبة؛ حيث أن محلل البيانات منفصل عن نظام LimeWire "الرئيسي". ويعد تحميل البيانات المسجلة في قاعدة بيانات MySQL عملية اضافية اختيارية منفصلة .

٥ / ٢ / ٢ مكونات النظام

يتكون نظامنا من الوحدات التالية: IR + لتحسين البحث. نظام LimeWire مع بعض التعديلات؛ وحدات مسجل البيانات، الحامل والمحلل.

التعديلات التي أدخلت على نظام LimeWire الأساسي ضئيلة للغاية، والتي ينبغي أن تبسط دمج التحسينات لدينا وفقا للإصدارات المستقبلية من LimeWire.

/ ٢ / ١ / ٢ / IR +

تستخدم هذه الوحدة البيانات الوصفية المرتبطة بالملف وتستخدم تقنيات استرجاع المعلومات لترتيب النتائج وإزالة غموضها، وبالتالي تحسين جودة النتائج لعملية بحث معينة. وهي تنفذ وظائف الترتيب الإضافية التالية:

١. تواتر المصطلح: النتيجة التي يحتل واصفها معظم مصطلحات الاستعلام في المرتبة الأعلى.
٢. الكسر: يتم تصنيف النتيجة التي يغطي واصفها أعلى جزء من مصطلحات الاستعلام في المرتبة الأعلى.
٣. تشابه جيب التمام: يتم ترتيب النتيجة بناءً على درجة تشابه جيب التمام بين الاستعلام واصفه.
٤. TF / IDF: هذا يشبه تشابه جيب التمام، ولكن مع كل مصطلح مرجح بناءً على تردد المستند. تتطلب وظيفة التصنيف هذه تقديرًا لتكرار المستند العام لكل مصطلح محسوب باستخدام مستودع التخزين المحلي المشترك.
٥. حجم المجموعة: يتم تجميع النتائج التي تشير إلى نفس الملف وتكون درجة التصنيف هي حجم المجموعة. يتم تنفيذ وظيفة التصنيف هذه في معظم أنظمة تبادل الملفات P2P، بما في ذلك LimeWire.

هذه القائمة من وظائف الترتيب ليست حصرية. تم تصميم مكون IR + بحيث يكون تنفيذ وظائف الترتيب الإضافية أمرًا سهلاً.

/ ٢ / ٢ / ٢ تسجيل استعلام نظام LimeWire

يتدفق بروتوكول Gnutella الأصلي على استعلامات الشبكة، مما يحد من قابلية التوسع. وللتعامل مع شكلة القابلية للتوسعة، يتم استخدام أجهزة فائقة النظائر. جهاز Ultrapeer هو عبارة عن نظير يُعتبر موثوقًا بدرجة عالية وقادرًا على التعامل مع عبء عمل Gnutella بشكل أكبر. وتتيح مجموعة الألياف الفائقة وجود نظام أكثر ثنائية، مع وحدات فائقة المرونة وعقد شجرية في بنية الهيكل. تتدفق Ultrapeers إلى استعلامات بعضها البعض، وتتصل العقد فقط بالـ Superpeers (وليس بالعقد الأخرى). يمكن النظر إلى الأجهزة فائقة الدقة على أنها "وكلاء" لعقد شجرية في شبكة Ultrapeers.

تتفاعل العقد مع الحروف الفائقة للتحكم في مقدار الاستعلامات الواردة بواسطة بروتوكول توجيه الاستعلام (QRP). يحدد QRP أن العقد تنشئ جداول توجيه الاستعلام عن طريق تجزئة الكلمات الأساسية لجميع واصفات الملفات التي تقوم بمشاركتها وتخزين قيم التجزئة هذه في متجه بت. من خلال تبادل جداول التوجيه هذه مع النظراء، تعرف العقد ما يمكن أن يطابقها نظرائهم واتخاذ قرارات التوجيه بشكل مناسب. تُنشئ العقد الشجرية جداول التوجيه هذه، وترسلها إلى أجهزة الطباعة

الطرفية الخاصة بها. وهكذا، لا تقوم Ultrapeers بتوجيه سوى مجموعة فرعية من الاستعلامات الواردة، والتي من المحتمل أن يتم الرد عليها، إلى عقدها.
على وجه التحديد، تكون الاستعلامات ملتصقة - يتطابق الاستعلام مع واصف إذا تم تضمين جميع مصطلحات الاستعلام في الواصف. يتلقى جدول التوجيه مع المزيد من وحدات الـ بت، مزيدًا من الاستعلامات. في الممارسة العملية، يتم تعيين الـ بتات كملفات تضاف إلى مستودع التخزين المحلي المشترك لمساعدة العملاء الذين لديهم عدد قليل من الملفات المشتركة، لذلك، يكون لديهم ناقلات بت متفرقة للغاية.
هدفنا هو تسجيل جميع طلبات البحث التي يتم إصدارها في شبكة Gnutella للحصول على صورة دقيقة عما يبحث عنه مستخدمو أنظمة مشاركة الملفات من نظير إلى نظير. إحدى الطرق للقيام بذلك هي مشاركة أكبر عدد ممكن من الملفات. هذا الخيار غير واقعي، بالنظر إلى حجم متجه البتات (٦٤ كيلو بايت). خيار آخر هو جعل عقدة LimeWire الخاصة بنا نظيرًا فائقًا. هذا الخيار أيضًا، ليس عمليًا، لأنه يتطلب إما الحفاظ على عقدة Gnutella ذات النطاق الترددي العالي لفترة طويلة من الزمن، وبالتالي يتم انتخابه كطرف فائق، أو تعديل برنامج LimeWire للسماح للعميل بالتنكر كآلة فائقة. الأول صعب لأنه يتطلب موارد مخصصة طويلة الأجل، دون أي ضمانات بأن الناتج سوف يصبح فائق التطرف. هذا الأخير، الذي يتنكر كأنه فائق، أمر ممكن، ولكنه يدخل في حالة عدم الاستقرار في الشبكة إذا كان فائق الكثافة أي ينضم ويترك.
للحصول على جميع الاستفسارات الواردة لتحليلنا، نختار خيارًا أسهل. لقد قمنا بتعيين جميع وحدات البت في ناقل البيانات للإدعاء الخاطئ بأننا نشارك الملفات التي تحتوي على جميع الكلمات الرئيسية التي يبحث عنها المستخدمون. هذا يتسبب في تعيين ناقل البيانات ذو الاتجاه الفائق مع كل الموجهات أيضًا، مما يؤدي إلى توجيه كل حركة مرور الشبكة إلى ذلك النوع من القطع المفرطة، ثم، بالطبع، بالنسبة لنا. قد يتسبب هذا في حدوث بعض الفيضانات في الشبكة ولكنه ضروري في إنتاج مجموعة بيانات غير متحيزة.

٥ / ٢ / ٢ / ٣ وحدات مسجل البيانات، الحامل والمحلل

هذا المكون مسؤول عن تسجيل وتحليل جميع أنواع الاستعلامات المختلفة مع النتائج التي تم الحصول عليها من البحث. يحتوي على المكونات الفرعية التالية:
ملفات السجل: تحتفظ IR-P2P بثلاثة أنواع من ملفات السجل. بخلاف تخزين الاستعلامات الواردة في ملف سجل الاستعلام الوارد ، يتيح IR-P2P أيضًا للمستخدم تسجيل الاستعلامات الصادرة وكذلك نتائج الاستعلام في ملف سجل آخر. يجري العمل أيضًا لتسهيل تخزين البيانات التي يتم استردادها من استعراض مستودعات

أقرانهم المشتركة. يتيح LimeWire للمستخدمين عرض محتويات النظير من خلال مرفق "استعراض المضيف". سيستخدم IR-P2P هذا المرفق لجمع البيانات على مضيف، ثم يستخدم هذه البيانات لإجراء عمليات بحث أخرى وإجراءات استعراض المضيف الأخرى. وبالتالي، فإنه سيتم إنشاء لقطة للبيانات المشتركة من قبل جميع المضيفين.

١. مسجل البيانات: يقرأ مسجل البيانات من ملفات السجل ويكتبها. يتم استدعاء وظائف مسجل البيانات من داخل رمز LimeWire لتسجيل البيانات إلى ملفات السجل المناسبة أو عرض السجل من خلال واجهة LimeWire.
٢. حامل البيانات: يقوم الحامل بقراءة ملفات السجل ويخزن البيانات في جداول MySQL المقابلة. لا يوجد لديه تفاعل مع كود LimeWire. يتم التحميل كعملية دفعية إذا أراد المستخدم استخدام الجداول.
٣. محلل البيانات: يسمح محلل البيانات للمستخدم بإجراء تحليلات متنوعة من خلال واجهة مستخدم بسيطة حيث يمكن إجراء التحديدات. سيكون المستخدم قادرًا على كتابة استعلامات SQL إلى قاعدة البيانات وحفظ النتائج المستردة أو الاختيار من التحليل الذي تم تنفيذه بالفعل. التحليلات التي تم تنفيذها بالفعل تخص رسائل الاستعلام الواردة. يتضمن ذلك: استرداد متوسط طول الاستعلام، وتوزيع طول الاستعلام، وتوزيع شعبية الاستعلامات، وتصنيف الاستعلام حسب نوع الملف (الصوت / الفيديو / البرنامج) الذي يريده المستخدم، وتحليل الارتباط. هذه الوحدة أيضًا لا تتفاعل مع كود LimeWire. يتم إجراء التحليلات على البيانات على ملفات السجل أو يتم إجراء الاستعلام على الجداول وفقًا لتفضيلات المستخدم.

٥ / ٣ محددات اداة البحث IR-P2P

يقتصر إصدارنا الحالي من IR-P2P على إمكانات تسجيل البيانات نظرًا لخصائص بروتوكول Gnutella. لا تسمح لنا هذه التفاصيل بتجميع "بيانات جلسة" محددة على المستخدمين، مما يمنع تحليل الجلسة (على سبيل المثال، كيف يقوم المستخدم بتعديل استفساراته لتحسين النتائج التي تلقاها)، يتم توجيه استعلام Gnutella على أساس قفزة. يتم توجيه الاستعلامات من نظير إلى جيرانه، الذين لا يتم إعطاء معلومات حول المكان الذي حصل فيه النظير على الاستعلام منه في المقام الأول. وبالتالي، يجب أن تتبع ردود الاستعلام (التي تحتوي على عنوان IP لعقدة الرد) مسارًا عكسيًا للوصول إلى منشئ الاستعلام.

من المحتمل أن يكون المسار العكسي ميزة أمان، مما يزيد من إخفاء هوية المستخدمين. هناك قيد آخر يتمثل في أننا لم نتمكن من الحصول على الوقت الذي تم فيه إرسال الاستعلام من قبل المستخدم. هذا وحقيقة أننا لا نستطيع تحديد منشئي

طلبات البحث تمنعنا من إجراء تحليل الجلسة الذي قد يساعدنا على دراسة سلوك المستخدم.

٥ / ٤ تحليل سجل الاستعلام

نقوم هنا بتحليل سمات سجل الاستعلام الوارد وعرض النتائج التي توصلنا إليها. ويهدف هذا التحلل فقط إلى إظهار أي نوع من التحليلات الممكنة مع البيانات التي تم جمعها بواسطة IR-P2P.

٥ / ٤ / ١ محرك بحث LimeWire

يسمح LimeWire للمستخدمين بمشاركة الملفات من أي نوع، مثل .mp3 و .avis و .jpg و tiffs وما إلى ذلك. إنه قادر على البحث المتزامن المتعدد والمتوفر بعدة لغات مختلفة. كما أنه يتيح للمستخدمين إجراء العديد من أنواع البحث المختلفة. بخلاف الاستعلامات العادية القائمة على الكلمات الرئيسية، يمكن للمستخدمين أيضًا البحث عن البيانات المشتركة على عنوان IP محدد. حتى بالنسبة لاستعلامات الكلمات الرئيسية، يمكن للمستخدم الاستعلام باستخدام سلسلة الاستعلام فقط أو يمكنه تحديد سمات أخرى مثل النوع أو الألبوم أو الفنان. يُطلق على هذا النوع من الاستعلامات "استعلام غني". في البحث المستند إلى الكلمات الرئيسية، يمكن للمستخدم إما البحث غير المقيد الذي يبحث عن الملفات من أي نوع أو البحث المقيد عن نوع معين من الملفات مثل الصوت أو الفيديو أو البرنامج الخ

في تحليلنا، قمنا بتقييد أنفسنا على طلبات البحث القائمة على الكلمات الرئيسية والتي ليست استعلامات غنية. نحن أيضًا نحصر أنفسنا في استعلامات اللغة العربية فقط (تتيح LimeWire الاستعلامات بعدة لغات مختلفة). كما سنظهر، لا ينبغي أن تضر هذه القيود بشكل كبير بعمومية نتائجنا، حيث أن معظم طلبات البحث تعتمد على الكلمات الرئيسية واستفسارات عربية.

٥ / ٤ / ٢ سجل استعلام IR-P2P الوارد

حافظنا على تشغيل IR-P2P طوال اليوم لجمع الاستعلامات الواردة يوم الجمعة، ٦ مايو ٢٠٢٠. خلال عطلة نهاية الأسبوع، كان هناك عدد أقل من زيارات العمل وعدد أكبر من الزيارات الترفيهية على الشبكة، مقارنةً بحالة أيام الأسبوع. نظرًا لأنه يمكن استخدام Gnutella بشكل أساسي لأغراض الترفيه، يجب أن نتمكن من تقديم مجموعة من الاستعلامات تمثيلية معقولة في عطلة نهاية الأسبوع. جمعنا البيانات في ٢٠ ملف سجل لكل منها ١٠٠ ميجابايت. كل سطر في ملف السجل هو طلب استعلام. يتم تخزين السمات التالية لكل استعلام: سلسلة الاستعلام الأصلية، وقت استلام الاستعلام، والقيم للإشارة إلى ما إذا كان البحث مقيدًا أم غير مقيد ونوع الملف الذي يبحث عنه المستخدم (مثل الصوت والفيديو، برنامج، وثيقة، صورة).

نقوم بمعالجة جميع سلاسل الاستعلام مسبقًا بتجاهل أي اختلافات في الحالة وإزالة كلمات التوقف واستبدال أي علامات ترقيم بمسافة بيضاء وضغط المسافة البيضاء على مسافات مفردة. علاوة على ذلك، لاحظنا أن بعض الاستعلامات تتضمن سلاسل غير معروفة. نعتقد أن معظمها عبارة عن استفسارات غير العربية، والتي تمت إزالتها من التحليل. هذا يشكل حوالي ٢٥ ٪ من سجل الاستعلام.
بعد المعالجة المسبقة، هناك إجمالي ٧٧٥٦٠٥ استعلامات و ٦٥٪ (٤٩٨,١٣٠) منها استعلامات مميزة دون النظر في النسخ المتماثلة.

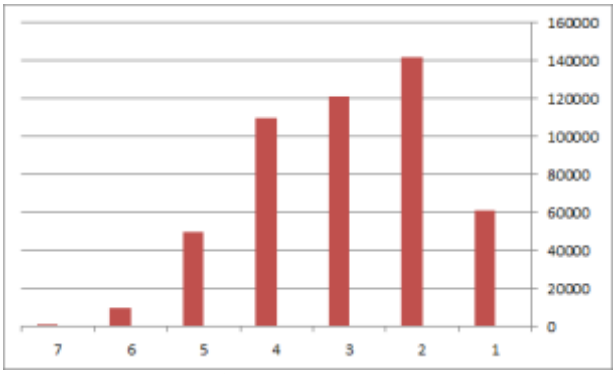
المبحث الخامس: رؤى ونتائج الدراسة.

كما هو موضح في الجدول (٥-١)، يبلغ متوسط عدد المصطلحات لكل استعلام ٢,٩٤، مما يؤكد أن الاستعلامات تميل إلى أن تكون قصيرة بشكل عام. تم الإبلاغ عن متوسط طول استعلام مماثل في دراسة سبينك وآخرون (٢٠٠٢)[٥٤]، مما يشير إلى وجود اتجاه استعلام قصير في كل من نظم مشاركة الملفات من نظير إلى نظير ومحركات البحث على الويب المركزية.

جدول (٥-١): يوضح احصائيات طول وتكرار الاستعلامات

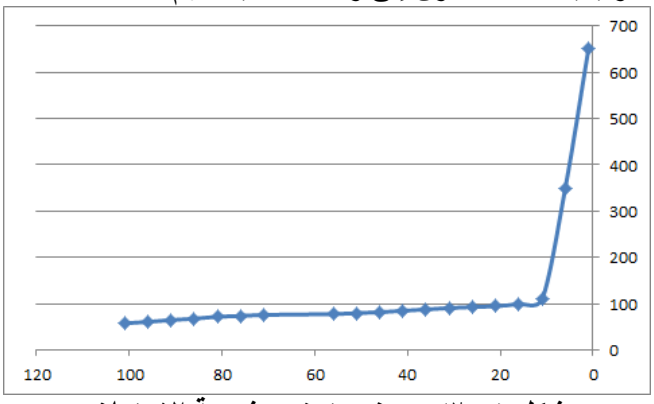
	طول الاستعلام	تكرار الاستعلام
الحد الاقصى	١٢	٦٢٣
الحد الادنى	١	١
المتوسط	٢,٩٤	١,٥٦
الانحراف المعياري	١,٢٧	٢,٢١

إلى جانب ذلك، فإننا نعرض توزيع طول الاستعلام كما موضح في الشكل (٥-٢). ويتم النظر فقط في الاستعلامات الفريدة في هذه الحالة. تحتوي معظم الاستعلامات على ١ إلى ٦ شروط استعلام. ولا يوجد سوى استعلام واحد يحتوي على ١٠ أو ١١ أو ١٢ مصطلحًا و ٣ طلبات بحث ذات طول ٩. ٧٥٪ من طلبات البحث تحتوي على ٢ إلى ٤ كلمات رئيسية.



شكل (٥-١): وضح توزيع طول الاستعلام

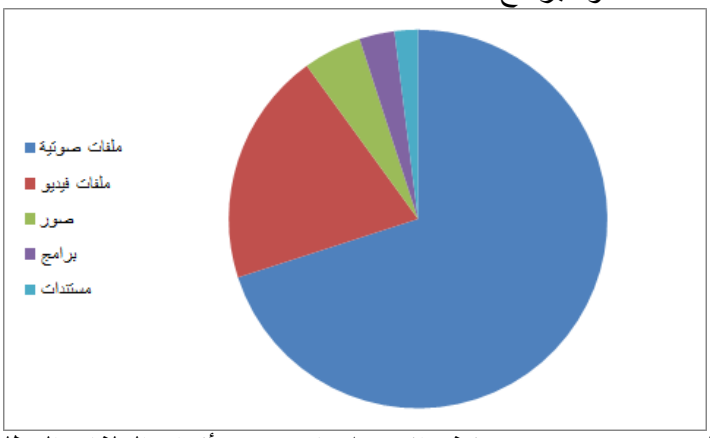
نقوم بمقارنة ترددات الاستعلامات الفريدة بترتيب ترددتها كما هو موضح في الشكل رقم (٥-١). فالاستعلام الذي تم تصنيفه في المرتبة الأولى هو الاستعلام الأكثر شيوعًا (التكرار) والاستعلام الأخرى هي الأكثر شيوعًا وما إلى ذلك. يتم التعامل مع الاستعلامات التي لها نفس شروط الاستعلام تمامًا كنسخ متماثلة لاستعلام واحد، حتى إذا قام العملاء بكتابة مصطلحات الاستعلام بترتيب مختلف. نلاحظ أن شعبية الاستعلام تتبع توزيع Zipflike (توزيع وفق قانون زيف). على عكس حقيقة أن تكرار الاستعلام الأكثر شيوعًا هو ٦٢٣، فإن حوالي ٧٥٪ من جميع الاستعلامات تحدث مرة واحدة فقط طوال اليوم، مما يدل على أن اهتمامات مستخدمي Gnutella متنوعة تمامًا أو يميل المستخدمون إلى وصف احتياجاتهم بشكل مختلف.



شكل (٥-٢): يوضح توزيع شعبية الاستعلام

قمنا بتصنيف جميع الاستعلامات المسجّلة استنادًا إلى الأنواع المطلوبة المضمنة في رسالة طلب الاستعلام. كما هو مبين في الشكل (٥-٣)، لا يحتوي جزء كبير من الاستعلامات على قيود على أنواع الملفات المطلوبة، والتي تمثلها وسيلة الإيضاح

"الكل". في الحالات التي يتم فيها تعريف واحد أو عدة أنواع من قبل العملاء إلى جانب الكلمات الرئيسية، فإن أكثر من ٧٠٪ من الاستعلامات مخصصة للملفات الصوتية، و ٢٠٪ مخصصة للفيديو، والاستعلامات الأخرى البالغة ١٠٪ مخصصة للصور والبرامج والمستندات. هذا هو نفسه كما توقعنا، لأن معظم العملاء في نظام مشاركة الملفات P2P يميلون إلى البحث عن الأغاني أو الأفلام بغرض الترفيه، بخلاف المستندات أو البرامج.



شكل (٥-٣): يوضح تصنيف الاستعلامات حسب أنواع الملفات المطلوبة

بعض امتدادات الملفات، avi، mp3 و dvd ، هي من بين أكثر شروط الاستعلام شيوعًا. السبب في اعتقادنا هو أنه بالنسبة للاستعلام الأصلي الذي يحتوي على كلمة رئيسية مثل أغنية.mp3، فإنه يصبح مصطلحين هما mp3 و mp3 بعد استبدال كل علامات الترقيم بمسافة بيضاء والتي يتم استخدامها لاكتشاف وفصل كلمتين رئيسيتين متجاورتين في مرحلة ما قبل المعالجة. نظرًا للكمية الكبيرة من طلبات البحث عن الموسيقى أو الأفلام، تحدث هذه المصطلحات عدة مرات وتُحسب كشروط استعلام شائعة.

من خلال تحليل نتائج النظام فإن هناك العديد من المجالات الرئيسية التي من المهم التركيز عليها اليوم لإنشاء أنظمة استرجاع المعلومات من نظير إلى نظير فعالة. تعتمد القائمة التالية على التصميم الحالي ونتائجه:

١. الجمع بين نقاط القوة للمؤشرات العمومية والمحلية وتطوير الخوارزميات لتحويل المحتوى المناسب بسهولة من واحد إلى الآخر على أساس المصطلح أو شعبية الاستعلام. فالنظم الحالية لا تتطور بشكل جيد لأنها تعتمد فقط على إما إغراق الشبكة بالاستعلامات أو لأنها تتطلب شكلاً من أشكال المعرفة العمومية.

٢. لا توجد بنية توفر أفضل حل لجميع مشكلات استرجاع المعلومات من نظير إلى نظير، وتنطبق أشكال معمارية مختلفة على مواقف مختلفة.

٣. على الرغم من أن خصائص قابلية التوسع الجيدة ملازمة لنموذج نظير إلى نظير، إلا أن النظم التي ترغب في دعم البحث على نطاق الويب تحتاج إلى التركيز على التوزيع الفعال لحملها على عدد كبير من أقرانها: مئات الآلاف إلى الملايين. أحد الأسباب المهمة لذلك هو أن الأقران غير متجانسين من حيث السعة والاتصال وأنهم ليسوا أجهزة خادم مخصصة: عليهم أداء العديد من المهام الأخرى أيضًا.

٤. التركيز على نتائج البحث بدلاً من الملفات. وهو ما يعني تحويل الانتباه إلى الشبكات التي توفر الوصول إلى الملفات الخارجية التي تؤكد مهمة البحث: جوهر استرجاع معلومات نظير إلى نظير.

٥. التحقيق وتحسين أداء نتائج البحث. من المهم تحقيق توازن جيد بين تقديم نتائج جديدة بما فيه الكفاية، وعدم فرض تكاليف على الشبكة لتحديث تلك النتائج. يجب أن يعتمد هذا أيضًا على وتيرة حدوث طفرة في الموارد المشار إليها.

٦. تحسين التعامل مع عدم التجانس بين الأقران في البحث على الويب. لدى عدد قليل من الأقران الكثير من الوثائق التي يقدمها، في حين أن العديد من أقرانهم لديهم مجموعات أصغر بكثير. غالبًا ما يتم تخصيص هذه المجموعات الأصغر لموضوعات محددة مما يجعلها مناسبة لاستفسارات أكثر تحديدًا.

٧. تطبيق التجميع على مختلف المستويات لأنه يسهل عملية إنشاء أوصاف المورد وتوجيه الاستعلام ، مما يؤدي إلى تقليل زمن الوصول.

٨. تحسين كل من طبولوجيا وتوجيه الاستعلام، وخاصة لتجنب وتوجيه النقاط الساخنة في الشبكات. يدعم الهيكل الجيد كلاً من الفعالية والكفاءة، من خلال تمكين الاستعلام للوصول إلى نظير مستهدف ذي صلة في خطوات قليلة.

٩. التركيز على الدقة على الاستدعاء. قد يتطلب تحقيق استدعاء ١٠٠ بالمائة في نظم نظير إلى نظير البحث في جميع المؤشرات وهو مكلف للغاية. كما أنه ليس ضروريًا إذا كانت جودة نتائج البحث التي تم إرجاعها عالية بدرجة كافية. على الرغم من أن هذا يتطلب تقنيات أفضل نتيجة الانصهار. أدرك أنه في بحث الويب، لا يتصفح معظم المستخدمين الصفحة الأولى، بل يشاركون في إعادة صياغة الاستعلام.

١٠. تطوير آليات توزيع التغذية المرتدة في الوقت الفعلي. من الناحية المثالية، تتحسن جودة نتائج البحث باستمرار استنادًا إلى ملاحظات المستخدم كما هو شائع في

محركات البحث المركزية. يمكن استكشاف الاتجاه الناشئ المتمثل في اقتران هذا بالشبكات الاجتماعية.

١١.إنشاء عدد من مجموعات الاختبار القياسية الكبيرة التي تنطبق على الأنواع المختلفة لشبكات استرداد معلومات نظير إلى نظير.

الخاتمة والتوصيات

قدمنا في هذا البحث نظامًا لمشاركة ملفات P2P باسم IR-P2P مبني على نظام Gnutella في Limewire لمراقبة استعلامات المستخدمين وجمعها، من أجل الكشف عن العالم الحقيقي للاستعلامات في شبكة Gnutella. يهدف هذا العمل إلى تلبية الحاجة إلى أدوات البحث والبيانات الخاصة باسترجاع الملفات في مجال المشاركة نظير إلى نظير. قدمنا أيضًا ملاحظات وتحليلات لما يقرب من مليون استعلام وارد تم جمعها في السجل. تظهر النتائج الإحصائية أن الاستعلامات تميل إلى أن تكون قصيرة. يحتوي ٧٥٪ من طلبات البحث على ٢ إلى ٤ كلمات رئيسية. تتبع شعبية الاستعلام توزيعًا يشبه توزيع زيف الاحصائي ومعظم الاستعلامات مخصصة للموسيقى والأفلام. ترتبط الكثير من مصطلحات الاستعلام ارتباطًا وثيقًا بالجانب الزمني، لذلك قد تتغير شعبيتها بشكل كبير مع مرور الوقت.

سيتم تنفيذ تحليلات أخرى مختلفة للاستعلامات الواردة، وسيتم إجراء تحليلات جديدة على الاستفسارات الصادرة والنتائج التي يتم إرجاعها. من بين أشياء أخرى في خط الاستعلامات، هناك تسجيل وتحليل لاستعلامات "مضيف الاستعراض" التي ستوفر نظرة ثاقبة جيدة لمشاركة البيانات للمستخدمين.

بالنسبة للاستعلامات الواردة، يمكن استخدام تقنية استخراج البيانات (التنقيب عن قاعدة الاقتران) للكشف عن الارتباط بين مصطلحات الاستعلام، والعلاقة بين قيم سمات الاستعلامات الغنية، وكذلك الارتباط بين مصطلح الاستعلام والسمة - القيمة. هذه المعلومات، إذا كانت متوفرة، مفيدة للغاية، حيث يمكن استخدامها لتحسين بروتوكول توجيه الاستعلام لأنظمة النظراء.

يمكن ايضا جمع النتائج التي يتم إرجاعها بواسطة النظام لكل استعلام مستخدم. تتمثل إحدى الطرق في استخدام الاستعلامات الواردة المسجلة باعتبارها استعلامات العقدة الصادرة وتسجيل جميع النتائج التي تم إرجاعها. من الممكن استخدام الاستعلامات التي يتم إنشاؤها تلقائيًا. قد تعطينا مجموعة النتائج التي تم إرجاعها لكل استعلام بعض المعرفة حول مدى كفاءة البحث في شبكة نظير إلى نظير.

يمكن ايضا التحقق مما إذا كان هناك ارتباط بين المجموعة المشتركة للمستخدم ومجموعة الاستعلامات التي أصدرها. تخميننا هو أن المستخدمين يبحثون عادة عن

ملفات مشابهة لتلك التي يشاركونها. ولكن الامر في حاجة الى التحقق تجريبيا من هذا الامر.

مراجع الدراسة:

1 K. Aberer, F. Klemm, M. Rajman, and J. Wu. (2004). "An Architecture for Peer-to-Peer Information Retrieval". Proc. of the 7th Annual Intl. ACM SIGIR Conf. Wrkshp on Peer-toPeer Information Retrieval.

2 O. Babaoglu, H. Meling, and A. Montresor. (2002). "Anthill: A Framework for the Development of Agent-based Peer-to Peer Systems". Proc. of the 22nd Intl. Conf. on Distributed Computing Systems (ICDCS'02).

3 C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. (1999). "Analysis of a Very Large Web Search Engine Query Log". SIGIR Forum, 33(1):6-12.

4 S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. (2004). Hourly Analysis of a Very Large Topically Categorized Web Query Log. SIGIR'04, 321-328, 2004.

5 D.Zeinalipour-Yazti, & T. Folias. (2002). A Quantitative Analysis of the Gnutella Network Traffic. TR-CS-89, Dept. of Computer Science, Univ. of California, Riverside.

6 Du, A. and Callan, J. (1998). "Probing a collection to discover its language model". Tech. Rep. UM-CS-1998-029, University of Massachusetts, Amherst, MA, US.

7 Lv, Q., Cao, P., Cohen, E., Li, K., and Shenker, S. (2002). "Search and replication in unstructured peer-to-peer networks'. In Proceedings of ICS. ACM, New York, NY, US, 84-95.

8 Skobeltsyn, G., Luu, T., Podnar š arko, I., Rajman, M., and Aberer, K. (2009). "Querydriven indexing for scalable peer-to-peer text retrieval". Future Generation Computer Systems 25, 89-99.

9 Lu, J. (2007). "Full-text federated search in peer-to-peer networks". Ph.D. thesis, Carnegie Mellon University.
10 Di Buccio, E., Masiero, I., and Melucci, M. (2009). "Improving information retrieval effectiveness in peer-to-peer networks through query piggybacking". In Proceedings of ECDL. 420-424.
11 Kirsch, S. T. (1997). "Document retrieval over networks wherein ranking and relevance scores are computed at the client for multiple database documents".
12 Luu, T., Klemm, F., Podnar, I., Rajman, M., and Aberer, K. (2006). "Alvis peers: A scalable full-text peer-to-peer retrieval engine". In Proceedings of P2PIR. ACM, New York, NY, US, 41-48.
13 Li, J., Loo, B. T., Joseph, L., Hellerstein, J. M., Karger, D. R., Morris, R., and Kaashoek, M. F. 2003. On the feasibility of peer-to-peer web indexing and search. In Proceedings of IPTPS. Lecture Notes in Computer Science, vol. 2735. Springer, 207-215.
14 Suel, T., Mathur, C., Wu, J.-w., Zhang, J., Delis, A., Kharrazi, M., Long, X., and Shanmugasundaram, K. 2003. Odissea: A peer-to-peer architecture for scalable web search and information retrieval. In Proceedings of WebDB. 67-72.
15 Stoica, I., Morris, R., Karger, D. R., Kaashoek, M. F., and Balakrishnan, H. (2001). "Chord: A scalable peer-to-peer lookup service for internet applications". SIGCOMM Computer Communication Review 31, 4, 149-160.
16 Yang, Y., Dunlap, R., Rexroad, M., and Cooper, B. F. (2006). "Performance of full text search in structured and unstructured peer-to-peer systems". In Proceedings of INFOCOM. 1-12

17 Reynolds, P. and Vahdat, A. (2003). "Efficient peer-to-peer keyword searching. In Proceedings of Middleware'. Lecture Notes in Computer Science, vol. 2672. Springer, 977-997.
18 Bloom, B. H. (1970). "Space/time trade-offs in hash coding with allowable errors". Communications of the ACM 13, 7 (July), 422-426.
19 Michel, S., Bender, M., Triantafillou, P., and Weikum, G. (2006). "Iqn routing: Integrating quality and novelty in p2p querying and ranking". In Proceedings of EDBT. Lecture Notes in Computer Science, vol. 3896. Springer, 149-166.
20 Song, W., Zeng, X., Hu, W., Chen, Y., Wang, C., and Cheng, F. (2010). "Resource search in peer-to-peer network based on power law distribution". In Proceedings of NSWCTC. 53-56.
21 Cuenca-Acuna, F. M., Martin, R. P., and Nguyen, T. D. (20030. "Planetp: Using gossiping to build content addressable peer-to-peer information sharing communities". In Proceedings of HPDC.
22 Zhang, J. and Suel, T. (20050. "Efficient query evaluation on large textual collections in a peer-to-peer environment". In Proceedings of P2P. IEEE Computer Society, Washington, DC, US, 225-233.
23 Balke, W.-T., Nejdl, W., Siberski, W., and Thaden, U. (2005). "Progressive distributed top-k retrieval in peer-to-peer networks". In Proceedings of ICDE. IEEE Computer Society, Washington, DC, USA, 174-185.
24 Skobeltsyn, G. and Aberer, K. (2006). "Distributed cache table: efficient query-driven processing of multi-term queries in p2p networks". In Proceedings of P2PIR. 33-40.
25 Tang, C. and Dwarkadas, S. (2004). "Hybrid global-local indexing for e cient peer-to-peer information retrieval". In Proceedings of NSDI.

26 Galanis, L., Wang, Y., Jeffery, S., and DeWitt, D. (2003). "Processing queries in a large peer-to-peer system". In Proceedings of CAiSE. Springer, Heidelberg, DE, 273-288.
27 Zeinalipour-Yazti, D., Kalogeraki, V., and Gunopulos, D. (20040. "Information retrieval techniques for peer-to-peer networks". Computing in Science and Engineering 6, 20-26.
28 Skobeltsyn, G. and Aberer, K. (2006). "Distributed cache table: efficient query-driven processing of multi-term queries in p2p networks". ibid. 33-40.
29 Bawa, M., Manku, G. S., and Raghavan, P. (2003). "Sets: Search enhanced by topic segmentation". In Proceedings of SIGIR. ACM, New York, NY, US, 306-313.
30 Akavipat, R., Wu, L.-S., Menczer, F., and Maguitman, A. G. (2006). "Emerging semantic communities in peer web search". In Proceedings of P2PIR. ACM, New York, NY, USA, 1-8.
31 Klampanos, I. and Jose, J. M. (2007). "An evaluation of a cluster-based architecture for peer-topeer information retrieval". In Proceedings of DEXA. 380-391.
32 Monnerat, L. and Amorim, C. (2009). "Peer-to-peer single hop distributed hash tables". In Proceedings of GLOBECOM 2009. 1-8.
33 Lv, Q., Cao, P., Cohen, E., Li, K., and Shenker, S. (20020. "Search and replication in unstructured peer-to-peer networks". In Proceedings of ICS. ACM, New York, NY, US, 84-95.
34 Kalogeraki, V., Gunopulos, D., and Zeinalipour-Yazti, D. (2002). "A local search mechanism for peer-to-peer networks". In Proceedings of CIKM. ACM, 300-307.

35 Adamic, L. A., Lukose, R. M., Puniyani, A. R., and Huberman, B. A. (20010. "Search in power-law networks". Physical Review E 64, 4 (Sept.), 046135-1-046135-8.
36 Yang, B. and Garcia-Molina, H. (20020. Efficient search in peer-to-peer networks". Proceedings of ICDS.
37 Tsoumakos, D. and Roussopoulos, N. (2003). "Adaptive probabilistic search for peer-to-peer networks". In Proceedings of P2P. IEEE Computer Society, 102-110.
38 Zhong, S., Chen, J., and Yang, Y. R. (2003). "Sprite: A simple, cheat-proof, credit-based system for mobile ad-hoc networks". In Proceedings of INFOCOM.
39 Li, C., Yu, B., and Sycara, K. (20090. "An incentive mechanism for message relaying in unstructured peer-to-peer systems'. Electronic Commerce Research and Applications 8, 6, 315-326.
40 Waterhouse, S., Doolin, D. M., Kan, G., and Faybishenko, Y. (20020. "Distributed search in p2p networks". IEEE Internet Computing 6, 1 (Jan/Feb), 68-72.
41 Suel, T., Mathur, C., Wu, J.-w., Zhang, J., Delis, A., Kharrazi, M., Long, X., and Shanmugasundaram, K. (2003). "Odissea: A peer-to-peer architecture for scalable web search and information retrieval". In Proceedings of WebDB. 67-72.
42 Bender, M., Michel, S., Triantafillou, P., Weikum, G., and Zimmer, C. (2005). "Minerva: Collaborative p2p search". In Proceedings of VLDB (Demos). 1263-1266.
43 Michel, S., Triantafillou, P., and Weikum, G. (2005). "Minerva infinity: A scalable efficient peer-to-peer search engine". In Proceedings of Middleware 2005. Springer, Heidelberg, DE, 60-81.
44 Luu, T., Klemm, F., Podnar, I., Rajman, M., and Aberer, K. (2006). "Alvis peers: A scalable full-text peer-to-peer retrieval

engine". In Proceedings of P2PIR. ACM, New York, NY, US, 41-48.

45 Rosenfeld, A., Goldman, C. V., Kaminka, G. A., and Kraus, S. (20090. "Phirst: A distributed architecture for p2p information retrieval". Information Systems 34, 2, 290-303.

46 Klampanos, I. A. and Jose, J. M. (2004). "An architecture for information retrieval over semicollaborating peer-to-peer networks". In Proceedings of SAC. ACM, New York, NY, US, 1078-1083.

47 Joseph, S. (2002). "Neurogrid: Semantically routing queries in peer-to-peer networks". In Proceedings of Networking Workshops. 202-214.

48 Galanis, L., Wang, Y., Jeffery, S., and DeWitt, D. (2003). "Processing queries in a large peer-to-peer system". In Proceedings of CAiSE. Springer, Heidelberg, DE, 273-288.

49 Triantafillou, P., Xiruhaki, C., Koubarakis, M., and Ntarmos, N. (2003). "Towards high performance peer-to-peer content & resource sharing systems". In Proceedings of CIDR. 120-132.

50 Callan, J., Lu, Z., and Croft, W. B. (1995). "Searching distributed collections with inference networks". In Proceedings of SIGIR. ACM Press, 21-28.

51 Bender, M., Michel, S., Triantafillou, P., and Weikum, G. (2007). "Design alternatives for large-scale web search: Alexander was great, aeneas a pioneer, and anakin has the force". Proceedings of LSDIR2007 Workshop.

52 Naicken, S., Livingston, B., Basu, A., Rodhetbhai, S., Wakeman, I., and Chalmers, D. (2007). "The state of peer-to-peer simulators and simulations". SIGCOMM Computer Communication Review 37, 2 (Apr.), 95-98.

53 Skobeltsyn, G., Luu, T., Podnar š arko, I., Rajman, M., and Aberer, K. (2009). "Querydriven indexing for scalable peer-

to-peer text retrieval". Future Generation Computer Systems 25, 89-99.

54 A. Spink, S. Ozmutlu, H. C. Ozmutlu, and B. J. Jansen. (2002). "U.S. Versus European Web Searching Trends". SIGIR Forum 36(2), 32-38.