# Gauging the Internet Doctor: Ranking Medical Claims based on Community Knowledge

V.G.Vinod Vydiswaran
University of Illinois
Urbana, IL
vgvinodv@illinois.edu

ChengXiang Zhai
University of Illinois
Urbana, IL
czhai@cs.uiuc.edu

Dan Roth
University of Illinois
Urbana, IL
danr@illinois.edu

## ABSTRACT

As more and more content is published and consumed online, it is imperative to know if an information nugget found on the Web is trustworthy or not. This is especially important for online medical information as it affects the most vulnerable group of users looking for medical help online. In this paper, we study the feasibility of automatically assessing the trustworthiness of a medical claim based on community knowledge, and propose techniques to assign a reliability score for an information nugget based on support over a community-generated collection. Specifically, we model the trustworthiness of a medical claim based on experiences shared by users in health forums and mailing lists. The proposed claim scores can be used to rank related claims on their relative trustworthiness. We further extend the notion of trustworthiness to a site (or equivalently, a database of claims from the site) and propose a scheme to rank sites based on aggregating the trust scores of claims from the site. Our experiments show that community knowledge can be exploited to help users distinguish reliable medical claims from unreliable ones. The proposed techniques can be applied to other domains where similar corpora are available.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation*; I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*NLP*

## General Terms

Design, Algorithms, Measurement, Experimentation

## Keywords

Trustworthiness, relation retrieval, forum credibility

## 1. PROBLEM MOTIVATION

With the advent of Web 2.0, it has become easier to publish, share, and consume content online. More and more information is sought over the Web, and the lack of control over what gets published online can lead to dissemination of unreliable and misleading information. This is especially worrisome in case of medical information, where quacks, alternative healers, and some pharmaceutical companies tout unproven remedies as miracle cures to unsuspecting patients. A "new" disease often triggers a mushrooming of medical sites, products, and unsolicited advice that can mislead the online audience. For instance, in October 2009, FDA [2] had to issue warnings to sites and certain well-known companies selling consumer products to rein in false and exaggerated advertisements about *Swine Flu* and its treatments.

Surveys show that more people across age groups now rely on online content for news, opinion, social networking, and other aspects of personal well-being, including health. A poll conducted in January–March 2009 [7] found that 62% of Americans got their information about HIV and AIDS through media sources, including news websites and the Internet, as compared to only 13% that received the information from their doctors. Other surveys [6] suggest that out of the 78% of Internet users who look for health information, 87% believe that the information they read online about health is reliable and only 20% validate the information by visiting authoritative websites such as CDC [1]. Given this reliance on online content, it is necessary to know if information from a site is trustworthy or not. Automatically labeling assertions as reliable or unreliable would help users distinguish between quacks and healers, between information spread by alternative medicine advocates and those approved by authoritative sources such as FDA and CDC.

However, verifying reliability of an information nugget and assigning an absolute trustworthiness score is challenging. This is especially true with medical information, since the effectiveness of particular drugs and procedures – and hence, the general opinion about a treatment regimen – varies widely with patients and cohorts. We postulate that although users must examine a claim in detail to ultimately determine its trustworthiness, we can prioritize the information to be examined based on predicted trustworthiness and raise an alert if the information seems suspicious.

We propose to rely on community experience to gauge reliability of an information nugget. Since thousands of patients share their experience with particular treatment methods on online forums, gleaning information from forums could potentially lead to strong indicators of treatment effectiveness. In this paper, we focus on finding relevant evidence to help a user validate a claim based on its support in a community-generated corpus. By promoting relevant supporting evi-


*KDD-DMH '11,* August 21, 2011, San Diego, California, USA.

dence, we believe the user can make a more informed decision about the trustworthiness of the claim, and hence avoid getting cheated by possibly malicious or spurious claims.

We formulate the problem of assigning trustworthiness to claims as a relation retrieval problem and address the following two questions: (i) Given a large text corpus of user generated content obtained from forums and mailing lists, and a database of claims expressed as binary relations (or tuples), how can we rank and score claims based on their verifiability (support) in the text corpus? (ii) Can we aggregate such information to say something holistic about the quality of the database itself, by ranking databases as more (or less) reliable, based on extent of support for each claim in the database? Note that we intentionally framed our research questions solely on text information in the community resources for two reasons: (i) Our focus is to see how useful text information alone is for predicting trustworthiness; this is a question that has never been studied in the existing work. Clearly, modeling user expertise or forum structure can be leveraged to further improve trustworthiness, but such extra information may not be available, whereas text information is always available. (ii) We want to focus on studying different features for improving the accuracy of the new text retrieval problem (i.e., relation retrieval problem) derived from the need for predicting trustworthiness, thus we did not pursue the goal of optimizing the prediction accuracy by using as much information as possible.

Automatic assessment of trustworthiness is important, yet fairly difficult challenge. This paper studies the feasibility of automatically predicting trustworthiness solely based on user-generated data, and makes the following contributions:

**1.** We propose the hypothesis that community knowledge can be leveraged to predict trustworthiness and develop an algorithmic framework to study if and how community knowledge can help predict trustworthiness of medical claims.
**2.** We define a novel relation retrieval problem and propose a suite of heuristic scoring methods for ranking both claims (from the same relation) and databases of claims in terms of trustworthiness, computed using multiple pieces of evidence from community-generated corpora.
**3.** We propose a novel evaluation scheme for predicting trustworthiness of medical claims. As evaluating such a task is difficult, we propose a scheme that adds challenging but wrong claims in a controlled way to a trusted database to make it less trusted, and measure the robustness of our framework to such perturbation.
**4.** We construct the first test set for evaluating trustworthiness of medical information on community portals. We have made public the full set of valid and invalid treatments used in the experiments to help other researchers evaluate potentially better methods on this dataset.[1]

We present the work in the medical domain, considering the relation between diseases and treatments as claims. The techniques are, however, applicable to other domains as well.

## 2. NOTIONS OF CLAIMS AND TRUST

In order to assign trustworthiness value to a piece of information, it is important to understand what kind of information it is, and what trustworthiness means for that piece of information. We postulate three kinds of claims:

[1]The complete dataset can be downloaded from `http://sifaka.cs.uiuc.edu/~vgvinodv/data.html`

**1. Physical claims:** These deal with information about things that have a physical existence and hence can be measured or verified accurately. This includes physical characteristics, such as height of mountains, distance between cities, authors of a book, etc. In such cases, trustworthiness of a piece of information often means how well it conforms to the actual value. The conformity can be measured by domain-specific distance measures. Some related work in this category include TruthFinder [27], that defines trustworthiness of online bookstores based on whether the list of authors or number of pages of a book have been correctly reported by the bookstore.

**2. Consensual claims:** These derive from information that is well-accepted by a group of experts or scientists working in a particular field. These are often not measurable directly, but most experts would agree on something being true, unless they find strong evidence to believe otherwise. Scientific concepts such as evolution, historic events such as conquests or the holocaust, etc. would fall under this category. Trustworthiness in such cases would quantify if some information conforms to this "well-accepted" belief.

**3. Perceived claims:** This class includes information that is based on preferences, perceptions, and opinions of an individual or a community. No unique, verifiable truth label can be assigned to the claim, and multiple values may be possible. Further, the label may change over short periods of time. Examples include political stand-points and product or movie reviews. Trustworthiness in such cases primarily involves modeling opinions and contrasting evidence that support and oppose the opinion. The information is preferred (gets higher "trust" score), if it conforms to one's perceptions and is not preferred, if it doesn't.

Under this scheme of classification, most controversies about trustworthiness arise in topics belonging to the category of **consensual claims**, since it assumes that a group of experts agree on the truth value of claims. Often, such expertise is not readily available to a common person. Hence, the person is left to infer trustworthiness of the claim based on the resources and evidence available at hand. Further adding to the lack of access to expert knowledge, many opinionated individuals could tout themselves as experts and present erroneous claims, giving supporting evidence of their own. This leads to alternative-but-partially-grounded truths to emerge and, in turn, motivates the need for some claim verification.

The medical domain is more susceptible to the limitations of the *consensual claim* verification. An ideal *medical claim verifier* would refer to the vast, trusted documentation of scientific literature or consult a team of experts in the field to validate a claim. This would, however, require significant effort in digitizing medical literature and assimilating clinical notes from numerous practitioners and patient databases.

## 3. MEDICAL TRUSTWORTHINESS USING COMMUNITY KNOWLEDGE

Medical practitioners believe that when dealing with different types of diseases, side-effects, and symptoms, it is quite difficult to exactly determine which treatments suit a particular individual or how much benefit a particular individual would get. Even after a drug is shown to be effective on a small sample of population and is approved, doctors learn about the effectiveness constantly through patient experiences as it is used more widely. Hence, most doctors,

health professionals, and regulatory bodies would find merit in a system that is able to assimilate information on effectiveness of drugs from shared experience of a large sample of patient population. Given the difficulties and the risks of relying on a single source that may represent special interests and biased view points, our underlying assumption is that a correct claim can be distinguished from an incorrect one by being supported by a diverse set of users. It is, hence, practical to analyze patient experiences to measure effectiveness of drugs. User-generated online resources, such as medical forums, discussion boards, and mailing lists, give us such a platform to model population knowledge. The main characteristics of a community knowledge-base are:

**(a)** It models the public at large. Typical forums attracts mostly patients and their concerned family members who share their experiences and ask for help and support from others in the community.
**(b)** It captures the sentiment around a particular treatment regimen, particularly dealing with contrastive opinion about various treatments.
**(c)** There is also a time component that captures if a drug effectiveness has varied over a period of time. Previous studies, such as [16], have shown how the user sentiments vary in tandem with initial FDA approval and later recall of now-infamous drugs such as *Vioxx*, *Tysabri*, and *Celebrex*.

Although not all information shared and published in health forums is reliable, we believe that the abundance of posts talking about effectiveness of particular drugs can still capture trustworthy information, since aggregated information over a large number of noisy signals have lower variances. This is especially pertinent for a relative comparison of different treatments for a particular disease. Rather than classifying posts into those that can be trusted and others that can't (or assigning scores to individual posts), we position our research to test whether the knowledge assimilated from community portals, forums, and discussion boards is reliable and trustworthy. This is in line with other works in research [19, 28] that tests validity of the "wisdom of crowds" conjecture that a community-driven knowledge base matches scientific expertise in quality. We propose techniques to effectively glean this knowledge of effectiveness of treatments, and help quantify our belief in its trustworthiness. Also, as described earlier, gauging effectiveness of a treatment based on sentiments expressed by patients is indeed a valid and relevant formulation in medical domain.

We cast the problem as a relation retrieval and ranking problem, where claims are scored based on the type of evidence we find from the corpus. We score each retrieved post on evidence of effectiveness of treatments, and then combine them into a normalized claim score. It is important to capture the relative strength of opinion rather than popularity as the score is used to *rank* treatments on effectiveness. Finally, the normalized scores of claims in a database are combined to score the database itself, allowing us to compare two databases of claims. We do normalization to overcome the concern of non-uniform coverage of different treatments, thereby avoiding bias against new or relatively unknown treatments.

## 3.1 Problem Formulation

To formalize the notation, (i) a claim $c$ is defined as an element of type $\langle \mathcal{E}_1, \mathcal{R}, \mathcal{E}_2 \rangle$, defined over two entity classes $\mathcal{E}_1$ and $\mathcal{E}_2$, and a relation $\mathcal{R}$ between the two entity classes. When considering a particular relation $r \in \mathcal{R}$, the entity classes are fixed and the domain of $c$ is uniquely defined.
(ii) A database of claims, $\mathcal{D}_r$, is defined as a set of relational tuples of type $\langle \mathcal{E}_1, \mathcal{E}_2 \rangle$, where the entities in each tuple are related by relation $r$.
(iii) Trustworthiness of a claim is a mapping $\phi: \langle \mathcal{E}_1, \mathcal{R}, \mathcal{E}_2 \rangle \rightarrow \mathbb{R}$, such that a highly trusted claim $c$ gets a high positive score $\phi(c) > 0$ and a highly untrusted claim gets a high negative score $\phi(c) < 0$. $\phi(c)$ is zero for a claim for which trustworthiness is unknown or cannot be determined.
(iv) Instead of formulating an absolute scoring function $\phi$ for a claim, a ranking function $\Omega$ can be defined over a set of claims $\mathcal{S}$, as $\Omega : \mathcal{S} \rightarrow \mathcal{O}$, where $\mathcal{O}$ is an ordering over elements of $\mathcal{S}$.
(v) A corpus $\mathcal{C}$ is defined as a collection of documents that is used to find supporting evidence for a claim. Hence, $\phi$ can be conditioned on $\mathcal{C}$.
(vi) Trustworthiness of a database of claims is a mapping $\psi : \mathcal{D} \rightarrow \mathbb{R}$, where $\psi$ is an aggregate function over $\phi$ and possibly other external parameters, such as $\mathcal{C}$.

Our objective is to test the hypothesis that the trustworthiness $\phi(c)$ of a claim $c$ can indeed be estimated using features from the corpus $\mathcal{C}$. This notion is then extended to define the trustworthiness $\psi(\mathcal{D})$ of a database $\mathcal{D}$ of claims.

As a specific instance of the formalism, we focus our work on the treatment relation in the medical domain. A treatment relation is expressed as a $\langle$`disease, treatment`$\rangle$ tuple, and a trustworthy instance of this tuple is one where the treatment has been approved to treat, prevent, or reduce the disease. Other examples from the medical domain include $\langle$`drug, side-effect`$\rangle$ and $\langle$`disease, symptoms`$\rangle$, that describe the relation of drug side-effects and disease diagnoses, respectively. Hence, in the current instantiation, $\mathcal{E}_1$ and $\mathcal{E}_2$ correspond to `diseases` and `treatments`, respectively, $\mathcal{R}$ is the relation "$\mathcal{E}_1$ is treated by $\mathcal{E}_2$", and $\mathcal{C}$ is a corpus consisting of messages in health forums and discussion boards, posted by a community of patients or people close to them.

Our goal is to assign scores to claims so that we can quantify the trustworthiness of treatments for a particular disease. Although different diseases may share treatments in some cases, it is not clear if treatments can be meaningfully compared across diseases. Hence, rather than learning absolute trustworthiness for all treatments, we learn a relative ranking of treatments for a disease. This ranking is based on whether the community believes a treatment is reliable and effective for a disease, ordered by their approval (or disapproval). This formulation also overcomes the concern of non-uniform coverage of diseases in the corpus. As we will show in Sec. 4.2, the non-uniform coverage of treatments for a particular disease is handled by normalizing the scores across all treatments mentioned for a disease.

# 4. SCORING CLAIMS BASED ON EVIDENCE

We split the problem into three steps:
**1.** Searching for relevant evidence documents that support the claim. The goal is to retrieve *all* occurrences of the treatment relation from the corpus $\mathcal{C}$.
**2.** Scoring individual evidence posts and claims by combining features from retrieved evidence via scoring functions $\phi$.
**3.** Aggregating the claim scores to compute trustworthiness score $\psi$ for a database of claims.

The following sections describe the three steps in detail. To simplify the explanation, we describe the steps in terms

of the ⟨`disease, treatment`⟩ relation instead of abstract claims, and the term "claim" refers to a ⟨`disease, treatment`⟩ relation tuple. Also, the terms *document* and *post* are used interchangeably.

## 4.1 Searching for evidence

Similar to the formulation in [24], the information need is modeled as a structured query consisting of three elements: the relation of interest, $\mathcal{R}$, and the two entities, $\mathcal{E}_1$ and $\mathcal{E}_2$, participating in the relation. The entities take specific roles in the relation, i.e., the entities are "typed" for a particular relation. Although this is a simplistic formulation of an arbitrary relation, we postulate that most direct relations are binary in nature and a long-distance relation could be modeled as a series of multiple binary relations. For example, an informational query that expects to find all diseases impacted by products from a pharmaceutical company could be modeled as two binary relations – ⟨`company, drug`⟩ and ⟨`drug, disease`⟩ – that are linked via the entity `drug`.

In this paper, we concentrate on a single relation of interest where the relation and the two entities partaking the relation are well-specified. Although the focus in this paper is on one specific relation, the suggested techniques are applicable to other relations and domains as well.

Our first task is to find from the corpus, *all* supporting forum posts that are relevant to the specified relation and provide evidence for the effectiveness of the treatment for the disease. To do so, we need to address the vocabulary mismatch problem, as posts might refer to the disease using a synonym, abbreviation, or a specialization of the disease. Typically, a domain-specific ontology such as MeSH [5] could be used to find valid synonyms for diseases and treatments. But, in order to make this component domain-independent, we did not use existing ontologies. Instead, we utilized the Wikipedia "redirect" link-graph structure. We searched for the Wikipedia page on the given disease or treatment name and collected the titles of all pages that point to and are pointed to from this page through the redirect links, to find relevant synonyms. For example, this way we could group "Chemo" and "Chemotherapy", and match up "impotence" and "ED" to "erectile dysfunction".

Similarly, the relation predicate also needs to be expanded. Instead of expecting all relation keywords to be given, we learned the set of words that can link the given entities. The relation words were extracted by first searching a Web corpus for common disease-treatment pairs (such as ⟨`diabetes, insulin`⟩) to retrieve matching sentences. These were parsed using a dependency parsing tool[2] [15] to find the most frequent patterns that connect the disease and treatment words. The head verbs from these patterns were considered as the relation verbs, and the verbs and their synonyms (from WordNet [17]) were added as relation keywords in the search query. We also added words that were *distributionally similar* to the relation words. The distributional similarity was computed over a large, independent, text corpus collected over the Web. Other works in literature [14, 20] show that words belonging to word-classes built in such an unsupervised fashion have high similarity and relatedness among them. Thus, our approach finds similar words in a general way. Although we need a few example tuples of the desired relation to seed the learning, these are usually easy to obtain. As a by-product of this approach, we get a relative frequency of occurrence (i.e., *importance*) of relation words, that is used in the next step.

Finally, the entity and relation components are combined to form a structured search query such that the retrieval system enforces the matching of both entities and at least one relation keyword in all retrieved documents. The retrieved documents were parsed using dependency parsing tools to confirm if the snippets also gave evidence for the presence of an explicit relation between the disease and treatment. This was then used the boost the rank of the document in the ranked list. It must be noted that this approach fails to recognize some relevant documents where the entities are not mentioned explicitly. This may lead to some forum posts getting ignored, especially those that are responses to an earlier post and the treatment or the disease are referred in the earlier post. Handling such cross-document co-reference resolution may be critical for some domains and is an interesting direction to extend this work.

## 4.2 Scoring claims based on evidence

Once all relevant posts are retrieved, the next task of scoring claims can be restricted over the retrieved set of posts.

### *4.2.1 Features for scoring functions*

The trustworthiness score is computed based on the following factors:

**1. Popularity of the treatment for the disease**, computed using number of posts that refer to the treatment in the context of the disease.
**2. Length of the post** indicates the type of information being shared. It was observed that most long posts are comparative in nature and useful to identify sentiment around the treatment being discussed.
**3. Number of opinionated words used,** contrasted to the length of the post.
**4. Positively and negatively opinionated words** indicate the sentiment towards the given treatment.
**5. Number of posts** that have a positive or negative orientation for a particular treatment.
**6. Number, type, and extent of subjective words** used in the post. Many repeated occurrences of subjective words in a post indicates a strong and consistent bias by the author rather than a single occurrence of subjective words near the treatment discussion and a neutral opinion elsewhere.

The subjectivity of retrieved posts is measured using a lexicon of over $7,000$ subjective words[3], that was used previously to analyze sentiments in text passages [26]. The lexicon consists of words and their polarity and subjectivity strengths. It is used to infer, for instance, that occurrence of the word "terrible" has a higher negative subjectivity than the word "hurt". Similarly, an adjective such as "dramatic" has high subjectivity, but is neither positive nor negative, and the polarity depends on the noun it modifies.

This feature set can be further extended with other features, notably based on forum post authorship. As we noted earlier, our focus in this work is to assess how well can we model trustworthiness based solely on textual features over forum posts. Hence, we modeled all users to have uniform expertise, which is also justified since users in the online health support groups tend to be patients rather than medical professionals.

[2]From `http://cogcomp.cs.illinois.edu/page/software`

[3]Downloaded from `http://www.cs.pitt.edu/mpqa/`

| Aggregation variants | Polarity variants | | | | Used to score |
|---|---|---|---|---|---|
| | **Opin** | **Subj** | **Scaled** | **Orient** | |
| | postOpin | postSubj | postScaled | postOrient | Posts |
| **Agg** | AggOpin | AggSubj | AggScaled | AggOrient | Claims |
| **Avg** | AvgOpin | AvgSubj | AvgScaled | AvgOrient | |
| **RankAvg** | RankAvgOpin | RankAvgSubj | RankAvgScaled | RankAvgOrient | |

**Table 1: Summary of scoring functions formulated**

#### 4.2.2 Formulating scoring functions

The next challenge is to formulate a scheme to combine the factors mentioned above into a scoring function. There are two key dimensions for the scoring function, viz., (a) how to include polarity information, and (b) how to aggregate the scores from multiple posts. We propose four variants to capture polarity: (i) using counts of polarity words alone, (ii) using strength of the polarity words used, (iii) scaling the counts with the strength of words, and (iv) categorizing the post as either positively or negatively oriented and using just the label. We call these as *Opin*, *Subj*, *Scaled*, and *Orient* variants, respectively, giving us four scores for each post.

Once a post is scored, the scores from all posts for a claim need to be combined to get a claim score. This can be done in one of two main ways: (i) averaging the scores, giving all posts equal weightage, or (ii) averaging the post scores using a differential weighting scheme based on the rank of the post in the retrieved ranked-list. These two variants are called *Avg* and *RankAvg*, respectively. Another option is to first aggregate the counts across all posts, and then compute the scoring function over these aggregated counts. This gives a third variant, *Agg*.

Table 1 summarizes the scoring variants explained above. These scoring functions are formalized below. Let us first define the notations used in the formulations:

- $w_i^+$ and $w_i^-$ show the number of positive and negative polarity opinion words in a post $p_i$.
- $s_i^+$ and $s_i^-$ show the number of positive or negative sentiment words in a post $p_i$, weighted by the subjective strength of the words used.
- $w^+$, $w^-$, $s^+$, and $s^-$ indicate similar measures, but the counts are accumulated over all relevant posts. E.g.

$$w^+ = \sum_{\text{posts } p_i} w_i^+ \tag{1}$$

- $p^+$ and $p^-$ show the number of positively or negatively oriented posts. A post $p_i$ contributes a count to $p^-$ if $w_i^+ < w_i^-$, else it contributes a count to $p^+$.
- $n$ denotes the number of relevant posts returned by a retrieval system.

#### 4.2.3 Scoring individual posts

Based on these parameters, the following functions were formulated to score posts:

**1. postOpin**: For a post $p_i$, this measure computes the average relative polarity of opinion expressed in the post as

$$\textbf{postOpin}(p_i) = \frac{w_i^+ - w_i^-}{w_i^+ + w_i^-} \tag{2}$$

**2. postSubj**: This computes the average relative subjectivity of opinion expressed in a post $p_i$, and is defined using the subjectivity features in addition to opinion features as

$$\textbf{postSubj}(p_i) = \frac{s_i^+ - s_i^-}{w_i^+ + w_i^-} \tag{3}$$

**3. postScaled**: Instead of measuring average subjectivity, we boost the relative polarity of opinion by the extent of subjectivity. For a post $p_i$, this measure is computed as

$$\textbf{postScaled}(p_i) = \left(\frac{w_i^+ - w_i^-}{w_i^+ + w_i^-}\right)^{\sigma} \tag{4}$$

$$\text{where } \sigma = \frac{s_i^+ + s_i^-}{w_i^+ + w_i^-}$$

**4. postOrient**: Depending on whether post $p_i$ is positively or negatively oriented, it can be scored as a binary function:

$$\textbf{postOrient}(p_i) = \begin{cases} -1 & \text{if } w_i^+ < w_i^- \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

#### 4.2.4 Scoring claims by aggregating over posts

Based on the basic post scoring functions mentioned above, the scores for a claim $c$ can be aggregated over all relevant posts in multiple ways (summarized in Table 1), as follows:

**1. Aggregating all relevant posts** into one "pseudo-post" and scoring this aggregated post: Instead of using post-specific counts, the aggregate counts are used, giving three variants, viz. *AggOpin*, *AggSubj*, and *AggScaled*. **AggOpin** is computed similar to **postOpin** (Eq. 2), except that it utilizes the aggregate counts $w^+$ and $w^-$ instead of post-specific counts $w_i^+$ and $w_i^-$. Similarly, **AggSubj** and **AggScaled** are defined using aggregate counts instead of post-specific counts in **postSubj** and **postScaled** formulations (Eq. 3 and Eq. 4), respectively.

For completeness, we also define **AggOrient** similar to **postOrient** (Eq. 5) over orientation of posts:

$$\textbf{AggOrient}(c) = \begin{cases} -1 & \text{if } p^+ < p^- \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

**2. Averaging the scores of individual posts** gives the next three variants: *AvgOpin*, *AvgSubj*, and *AvgScaled*. **AvgOpin** averages the **postOpin** scores (Eq. 2) over all relevant posts, giving each post a uniform weight.

$$\textbf{AvgOpin}(c) = \frac{1}{n} \cdot \sum_p \textbf{postOpin}(p) \tag{7}$$

**AvgSubj** and **AvgScaled** measures are similarly defined using **postSubj** and **postScaled** scores (Eq. 3 and Eq. 4), respectively. A fourth variant, **AvgOrient**, is computed over the orientation features $p^+$ and $p^-$, as

$$\textbf{AvgOrient}(c) = \frac{p^+ - p^-}{n} \tag{8}$$

**3. Combining post scores in a weighted average** gives the final set of variants: *RankAvgOpin*, *RankAvgSubj*, and *RankAvgScaled*. The weight is proportional to the rank of the post in the ranked list after the retrieval stage. Specifically, **RankAvgOpin** is computed as a weighted average of **postOpin** scores (Eq. 2). If $r(p)$ is the rank of the post $p$, then the measure is computed as

$$\textbf{RankAvgOpin}(c) = \frac{1}{n} \cdot \sum_p \textbf{postOpin}(p) \times \log(n+1-r(p)) \tag{9}$$

**RankAvgSubj** and **RankAvgScaled** measures are similarly defined using **postSubj** and **postScaled** scores (Eq. 3 and Eq. 4), respectively. These measure gives higher weight to posts near the top of the ranked list than to those at the bottom of the list. We evaluated many variations of weighting function, such as $1/r$, $1/\log(r+1)$, $\log(n+1-r(p))$, etc., and chose the formulation that performed the best.

Finally, the fourth variant, **RankAvgOrient**, uses the orientation features similar to **AvgOrient** (Eq. 8), but instead of a uniform weight of 1, each post $p$ gets a weight proportional to its rank $r(p)$.

$$\textbf{RankAvgOrient}(c) = \frac{1}{n} \cdot \sum_{p} \gamma(p) \times \log(n+1-r(p)) \quad (10)$$

$$\text{where} \quad \gamma(p_i) = \textbf{postOrient}(p_i) = \begin{cases} -1 & \text{if } w_i^+ < w_i^- \\ 1 & \text{otherwise} \end{cases}$$

## 4.3 Scoring database of claims

In addition to knowing which nuggets of information are reliable, it would be good to also extend the notion of trust to sources of information. A medical website that gives some information about a treatment, may also give information about other treatments for one or many diseases. A "claim verifier" system could extend the notion of trusted claims to rank websites based on the reliability of claims on that site.

Functionally, a website is modeled as a database of claims. Once the scores are computed for each claim, they are aggregated to compute the trustworthiness of the overall database. In order to compute an aggregate score of the claim database, we use the following weighted average measure:

$$\text{DBscore} = \frac{\sum_c n_c \times \text{score}(c)}{\sum_c n_c} \quad (11)$$

where the summation is over all claims in the database and $n_c$ is the number of times a claim $c$ appears in the database as a relation. For a claim, the weight is proportional to $n_c$.

# 5. EXPERIMENTS

The primary focus of our experiments was to investigate if community-knowledge based corpora can be used to measure reliability of an information nugget, and if so, what parameters affect the performance. In this section, we describe our experimental setup and data characteristics, and evaluate our approaches to answer the following research questions:

**1.** Can community knowledge in forums and discussion boards help identify reliable information?
**2**. Which parameters affect the technique's effectiveness?
**3**. Can we extend the trust modeling for claims to measure trustworthiness of a database of claims?

## 5.1 Test Set Construction

As we discussed in Sec. 3, medical procedures and treatments are often inexact and categorizing them as completely trustworthy or completely bogus may be incorrect. However, in order to evaluate the effectiveness of our techniques that assign trustworthiness rank to treatments, we needed to construct a high quality database of "gold-standard" ⟨`disease, treatment`⟩ pairs. We chose six widespread diseases or medical conditions, viz., AIDS, Arthritis (specifically, Osteoarthritis), Asthma, Cancer, COPD, and Impotence (specifically, male infertility). We preferred these diseases over others, such as Influenza, because these are some examples of chronic ailments that do not have definitive cures. So, we could find many potentially-competing treatments. Further, because of their chronic nature, there is a higher possibility of finding more discussions about these diseases in health forums.

| Disease | Treatments | |
|---|---|---|
| | **Approved** | **Alternate** |
| AIDS | Abcavir<br>Kivexa<br>Zidovudine<br>Tenofovir<br>Nevirapine | Acupuncture<br>Herbal medicines<br>Multivitamin<br>Tylenol<br>Selenium |
| Arthritis | Physical therapy<br>Exercise<br>Tylenol<br>Morphine<br>Knee brace | Acupuncture<br>Chondroitin<br>Gluosamine<br>Ginger rhizome<br>Selenium |
| Asthma | Salbutamol<br>Advair<br>Ventolin<br>Bronchodilator<br>Xolair | Atrovent<br>Serevent<br>Foradil<br>Ipratropium |
| Cancer | Surgery<br>Chemotherapy<br>Quercetin<br>Selenium<br>Glutathione | Essiac tea<br>Budwig diet<br>Gerson therapy<br>Homeopathy |
| COPD | Salbutamol<br>Smoking cessation<br>Spiriva<br>Oxygen<br>Surgery | Ipratopium<br>Atrovent<br>Apovent |
| Impotence | Testesterone<br>Implants<br>Viagra<br>Levitra<br>Cialis | Ginseng root<br>Naltrexone<br>Enzyte<br>Diet |

**Table 2: Sample list of diseases and treatments from the claim database**

We manually collected 106 treatments across these six diseases. These treatments consisted of names of drugs and drug classes, specific devices (such as "knee braces"), specialty procedures (such as "chemotherapy"), and lifestyle changes (such as "exercise"). The treatment information was collected from medical web-portals, such as WebMD [9] and Yahoo! Health [11], and from Wikipedia [10] articles for diseases. Some examples are shown in Table 2, column 2.

We augmented the database with 93 invalid treatments of two main types. The first type of invalid treatments were those that were disapproved (banned) or considered controversial (unscientific). Typical examples include alternative therapies, herbal medications, and mechanical devices that have been scientifically proven to be ineffective or potentially harmful. Some examples are shown in Table 2, column 3.

The second type of invalid treatments that we added were a list of common treatments, medications, supplements, or lifestyle changes, such as *Tylenol*, *Advil*, vitamins, exercise, etc., that have a high chance of co-occurring with certain ailments, but do not serve as prescribed treatments for specific diseases. An exception should be noted: for Arthritis, exercises such as walking and swimming are prescribed as lifestyle changes helpful in treating the disease.

Each disease-treatment pair, hence, had a label indicating if the treatment was valid, disapproved, or non-specific. Table 3 gives a distribution of number of valid and invalid treatments in our dataset. The counts for invalid treatments include both disapproved and non-specific treatments.

| Disease | Treatments considered | | |
|---|---|---|---|
| | Approved | Invalid | Total |
| AIDS | 22 | 10 | 32 |
| Arthritis | 21 | 19 | 40 |
| Asthma | 14 | 13 | 27 |
| Cancer | 15 | 25 | 40 |
| COPD | 12 | 10 | 22 |
| Impotence | 22 | 16 | 38 |
| Total | 106 | 93 | 199 |

**Table 3: Treatments considered in our claim dataset**

| Site | Corpus size | # posts |
|---|---|---|
| Yahoo! Health | 15 G | 12,520,438 |
| healthboards.com | 3.1 G | 2,730,667 |
| medhelp.com | 1.9 G | 1,621,677 |
| ehealth.com | 410 M | 438,499 |
| medicalconversation.com | 1.3 G | 432,185 |
| medicalreplies.com | 221 M | 187,580 |
| wrongdiagnosis.com | 145 M | 96,985 |
| doctorslounge.com | 80 M | 49,863 |
| mdhealth.com | 9.4 M | 5,402 |
| Total | 22.2 G | 18,083,296 |

**Table 4: Details of the collected corpus of medical forums and discussion boards**

Based on this gold-standard set of disease-treatment pairs, we constructed two test sets – `Skewed` and `Balanced`. For constructing the `Skewed` test set, we randomly sampled five valid treatments for every disease, and combined it with *all* the invalid treatments for that disease. We created 25 such combinations per disease, randomly picking the valid diseases each time. For `Balanced` test set, we built a similar test case with ten each of valid and invalid treatments per disease. Thus, we had $25 \times 6 = 150$ test cases in each set.

## 5.2 Corpus statistics

In order to collect a representative sample of community-generated corpus, we crawled a mix of popular online medical forums, including both large and small forums, over a period of two months in early 2009. We selected eight such medical forums. In addition, we got access to a large dump of Yahoo! Health groups messages. These messages are more personal interactions in small support groups, but are otherwise similar to the forums. The characteristics of the collected corpus is summarized in Table 4.

As part of cleaning the data, we removed all forum content except the text of the posts. We obfuscated user specific data for these experiments to treat all forum posts uniformly instead of distinguishing posts based on authorship. We understand that user information would be useful in other trust models but for the current setup, we chose to ignore that information. Similarly, we ignored the time information available with posts, though that would also be useful.

The individual posts were treated as separate documents, and sections of posts that were duplicates of an earlier post in the thread (which often happens someone responds to an earlier post in mailing lists) were removed. The cleaned posts were, on-an-average, about 150 to 200 words long.

We built a retrieval system over the corpus using the Lemur toolkit [8], and utilized the rich IndriQuery querying language to query the index. We first constructed a simple query with disease and treatment terms and their synonyms.

| Test set | System | MRR | P@5 | P@10 | MAP |
|---|---|---|---|---|---|
| `Skewed` | Random | 0.512 | 0.271 | 0.257 | 0.385 |
| | Popularity | 0.588 | 0.339 | 0.306 | 0.436 |
| | Ours | **0.681** | **0.369** | **0.319** | **0.469** |
| | Impr. (R) | 33.0%* | 36.2%* | 24.1%* | 21.8%* |
| | Impr. (P) | 15.8%* | 8.9%* | 4.3% | 7.6%* |
| `Balanced` | Random | 0.629 | 0.496 | 0.492 | 0.562 |
| | Popularity | 0.687 | **0.536** | 0.519 | 0.591 |
| | Ours | **0.703** | 0.524 | **0.525** | **0.596** |
| | Impr. (R) | 11.8%* | 5.6% | 6.7% | 6.1% |
| | Impr. (P) | 2.3% | −2.2% | 1.2% | 0.9% |

**Table 5: Comparison of ranking performance using multiple measures. Last two rows in each test set show relative improvement of our system (using `RankAvgSubj`) over (R)andom and (P)opularity baselines. * indicates statistical significance at $p = 0.05$.**

We extended the query by adding the relation words to the query, combined in such a way that the retrieval system enforces both entities and at least one relation word to match in all retrieved documents. Then, as explained in Sec. 4.1, we parsed the snippets to find instances of the treatment relation. Finally, the posts were re-ranked based on the occurrence of treatment relations in the result snippets.

## 5.3 Feasibility of scoring based on forum posts

We first wanted to validate if it is feasible to score treatments based on information available in forums. For this, we identified top 1000 posts relevant to a treatment. The posts were then scored on the extent to which they supported the treatments. These scores were then aggregated using one of the scoring functions defined in Sec. 4.2.

In our experiments, the query is a disease, and we rank the treatments based on their reliability score. We measure our performance using Mean Average Precision (MAP), Precision at 5 and 10 documents (P@5 and P@10, respectively), and Mean Reciprocal Rank (MRR). MAP is the arithmetic mean over all queries of average of precision values computed at ranks where relevant documents are retrieved. P@5 and P@10 measure the precision at top five and top ten retrieved results, respectively. MRR is the arithmetic mean of the multiplicative inverse of the rank of the top-most relevant result. For all measures, higher values are preferred.

Table 5 shows our performance (using the best scoring function) against two baselines – Random and Popularity. The random baseline just chooses the ranking of treatments at random. On the other hand, the popularity baseline ranks treatments based on number of posts returned by the retrieval system. The results show improvement over both baselines. In the `Skewed` dataset, we are able to improve the MRR and $P$@5 measures by 30% over the random baseline, which means that more relevant treatments are ranked higher in the list. We also perform significantly better than the popularity baseline on average, especially in the `Skewed` test set, which shows that adding additional features from the posts leads to better ranking. Since the number of valid and invalid treatments are equal in the `Balanced` dataset, improving over the proposed baselines is artificially harder.

### 5.3.1 *Effect of different scoring schemes*

We wanted to investigate how different schemes proposed in Sec. 4.2 perform against each other. We first measured the

| Aggregation variants | Polarity variants | | | |
|---|---|---|---|---|
| | **Opin** | **Subj** | **Scaled** | **Orient** |
| **Agg** | AggOpin 0.482 | AggSubj 0.547 | AggScaled 0.509 | AggOrient 0.354 |
| **Avg** | AvgOpin 0.524 | AvgSubj 0.536 | AvgScaled 0.359 | AvgOrient 0.481 |
| **RankAvg** | RankAvgOpin 0.569 | **RankAvgSubj 0.576** | RankAvgScaled 0.359 | RankAvgOrient 0.513 |

**Table 6: Variation of MAP with the different scoring schemes, for cancer treatments using Relation Retrieval approach. The baseline approach gives a MAP of** 0.3**.**

| Measure | Sentence | Passage | Document |
|---|---|---|---|
| MRR | 0.60 | 0.68 | 0.67 |
| MAP | 0.44 | 0.59 | 0.56 |

**Table 7: Variation of the size of context affects performance. Larger texts like passages and documents are better than only considering sentences.**

performance of a simplistic **scoring baseline**. We counted the number of posts returned for a treatment and used that as its score. Ranking cancer treatments with this baseline technique gave a MAP of 0.3. Then, to evaluate the different scoring functions, we ranked cancer treatments using all scoring variants defined in Sec. 4.2 (Table 1), and the comparison is shown in Table 6. We find that the *Scaled* variants have significantly lower MAP scores compared to other polarity variants. Further, the *Orient* variant (that treats each post as either positive or negative, without weighting) and the *Opin* variant (that only looks for opinion words and does not consider subjectivity) do not seem to be sufficient by themselves. *Subj* variants seem to consistently out-perform other variants, especially when used with rank-based scaling. The `RankAvgSubj` variant gives the best MAP scores. This indicates that both subjectivity and rank-based features seem to help in scoring trustworthiness of treatments.

### 5.3.2 *Effect of context used to compute scores*

Next, we wanted to investigate how the nature of context considered affects the overall performance. We varied the size of the context window used to find the subjective bias in a post. We considered three variants: a sentence level, a passage of three sentences surrounding the matched keywords, and the entire post. Table 7 summarizes the findings, when using the best scoring scheme, `RankAvgSubj`. It shows that passage level granularity seems to be better than considering only the sentence, indicating that larger contexts capture sentiments better. However, extending the context to the entire post degrades the performance slightly.

## 5.4 Assigning trust scores to claim databases

Our next goal is to investigate if the scoring and aggregating techniques can be used to also evaluate the trustworthiness of a site. As discussed earlier, we model a site as a database of claims. In order to evaluate how well we distinguish a trusted database from an untrusted one, we would have to ideally collect multiple databases of varying degrees of trusted information. Instead, we simulated such databases by gradually introducing errors in existing database of claims. We perturbed an existing claim database in a controlled manner, by replacing valid entries with invalid ones and observed if our computational method can pick up such errors. Assuming that we have a good scoring scheme for treatments, a database that has more valid treatments would score higher than another database with many erroneous, faulty, or irrelevant entries. By monitoring and varying the amount of noise through our simulations, we could evaluate the sensitivity of our scoring techniques in a controlled environment.

We started with a subset of the "gold-standard" dataset we had manually created and simulated degradation of claims by replacing valid treatments with invalid ones. So, we perturbed the claim database in two ways: (i) by introducing disapproved treatments per disease, and (ii) by introducing common (generic) treatments, such as paracetamol or *Advil*, that are not specific treatments for particular diseases, but may co-occur with many diseases.

For each disease, we started with 15 valid disease-treatment pairs, and gradually replaced random valid entries with invalid (disapproved or non-specific) treatments (cf. Table 3). This way, each modified set contained 15 disease-treatment pairs, but as more noise was added, the number of valid entries in the set kept reducing. The initial set of 15 disease-treatment pairs is chosen randomly from the valid entries; if a disease had fewer than 15 valid treatments, all were selected. Our choice of replacements is, however, not random but carefully controlled: we first add the invalid treatments into the claim database. Then, once those are exhausted, we add the non-specific treatments. After a swap is made, we compute the overall database score using Eq. 11, where individual claim scores are computed using the `RankAvgSubj` measure shown to be most effective (Sec. 5.3.1).

Fig. 1 shows the variation of database trustworthiness score with the addition of noise. We see that, as the database (of 15 treatments per disease) gets noisier, the aggregate trustworthiness score for the database reduces for all six diseases. The reduction is more significant in the initial 60%, when disapproved treatments are being added to the claim database. The change in the second phase (when non-specific treatments are added) is less significant.

We note that the plots for `Cancer` and `Asthma` have high scores at 0% degradation. This is consistent with the fact that these diseases have well-known and established treatments that work well and their "trustworthiness" scores are high. For the remaining four diseases, the treatments are not so well-established and the absolute values of the scores are smaller. However, in all six cases, with addition of invalid treatments, the overall score for the claim database reduces.

The case with the `Arthritis` plot is somewhat unique. The score seems to improve after 70% of the database gets

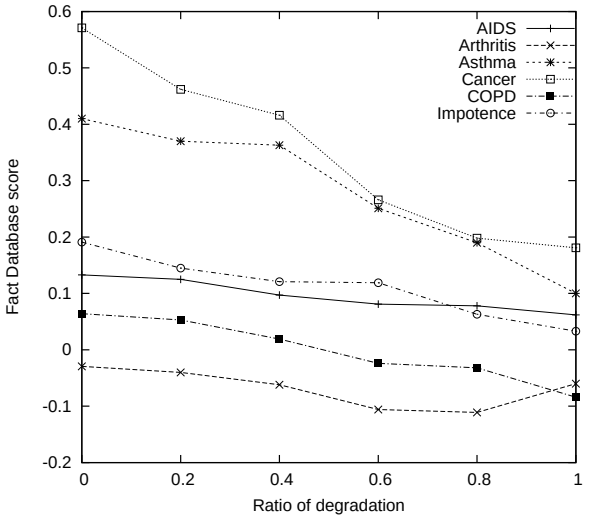


**Figure 1: Variation of database trustworthiness score with addition of noise, for all six diseases**

noisy. On deeper analysis, we found that the scores for the `Arthritis` treatments (even at 0% degradation level) are negative. However, the score assigned to non-specific treatments are usually close to zero (irrespective of the disease), since there is no strong bias in favor or against such treatments. So, in our simulation runs for `Arthritis`, once all the disapproved treatments (about ten in number) are included and the second group of non-specific treatments are added as noise, these near-zero scores increase the average trustworthiness score. This case shows that while the proposed scheme is effective in assigning higher scores to valid treatments than invalid treatments for a particular disease, when computing the overall database score across diseases, the proposed scoring scheme should be normalized to balance out any scoring bias for or against a particular disease.

# 6. RELATED WORK

There are several systems that allow users to search specifically for medical information. Gaudinat et al. [18] proposed a search engine for health documents. Hidola [3] is a personalized health and medical information search engine that allows searching personalized health information by selecting symptoms and answering questions in a medical questionnaire, rather than typing-in keywords. Commercial websites, such as WebMD [9], also allow searching diseases by clicking on symptoms. We believe that the our techniques could add value to the information from these sites by validating the treatment claims based on shared user experience.

On the reliability aspects of health information, governmental public health agencies, such as FDA and CDC, usually take active role in issuing warnings and thwarting rumors as part of their regulatory functions. Medical and scientific literature, such as [16], have used message boards to track important events in public healthcare, such as recall of drugs by governmental agencies. Some non-governmental agencies, such as Health on Net Foundation [4], offer services such as manually certifying websites as "trustworthy" based on a set of guidelines that depend on the site structure, its primary intent (commercial vs. informative), etc. They do not rate health forums or individual claims, and hence, could not be used for evaluation in this paper.

Our approach to ranking websites as a database of claims needs to be distinguished from other lines of work that model site score to be dynamically co-dependent on the claim score. TruthFinder [27], for example, computes the trustworthiness of a site based on the claims from the site, and vice versa. In one of our other works [25], we propose a content-driven trustworthiness framework that incorporates the trustworthiness of evidence for claims in the trust computation. In fact, the techniques proposed in this work to identify and score relevant posts could be used to instantiate a model to compute trustworthiness of medical claims and sources. Our main focus in this paper, however, is to verify if community knowledge can detect and measure trustworthiness of websites. Hence, in this paper, we model a database score as an aggregate function of claim scores alone.

There has been some work on extracting facts from the Web [13, 23] based on frequent patterns relating one entity to the other. The system proposed in [22] tries to extract facts by mining search query logs to identify relevant attributes for entities. We believe that analyzing forums for related entities follows similar principle of tapping into the "wisdom of crowds", except that instead of extracting entities, we formulate the problem as finding evidence in support of a given relation from a large text corpus. Our approach can be seen as adding supplementary trustworthiness information to the extracted facts.

Published literature includes a lot of research on sentiment and opinion analysis [21], primarily over product and movie reviews, that is relevant to the approach taken in this paper. Although deeper sentiment analysis would help in better understanding of text, such approaches expect the input text to be grammatically correct, which is not often the case in forum posts. Our focus is mainly to validate claims of treatments made in forums, rather that analyze the forum posts in detail, and as we have shown, this is indeed feasible with word-based measures and relation retrieval approaches. The work can be extended with stronger sentiment analysis.

Finally, the work may also be related to recommendation systems, since we can view the community-driven trustworthiness as the recommendation a community gives for a claim. Most work on recommendation systems [12] rank relations based on homophily and collaborative filtering. Our ranking scheme also involves components of homophily, since we prefer to rank a treatment higher if many others in the community prefer the treatment over other treatments.

# 7. CONCLUSION AND FUTURE WORK

In this paper, we have shown that it is feasible to score trustworthiness of claims based on the opinions expressed by millions of users in forums and message boards. We posed the problem as a relation retrieval problem and showed that by formulating the search as a structured query, we were able to garner more representative results which in turn helped compute better support for a claim. We evaluated this approach over six diseases and showed that the more "trusted" treatments get ranked over the untrusted ones. We extended the scheme to also score a database of claims, and demonstrated through simulations that as more noise was added to the claim database, the database trustworthiness score reduced. Our proposed techniques have been presented in the context of diseases and treatments in medical domain, but

can also be applied to relations from other domains where community-generated corpora are available.

Note that our focus was not to develop the *best* algorithm for predicting trustworthiness that would presumably require additional information about users and forum structure. Instead, our goal was to study if community-generated text can be reliably used to predict trustworthiness of claims. Our work clearly demonstrates the feasibility of using text information in the community resources. Adding features based on who wrote the post or when it was posted would further enrich the trustworthiness model. Note that even with the basic features we tested, our results show that we can already rank claims meaningfully based on their estimated trustworthiness. Indeed, our results are quite promising because we only used text information, so there is a great potential to further improve performance by leveraging user information, which is an interesting future work.

As a first step to test this general idea, we did not attempt to optimize many components in our scoring function including more accurate sentiment analysis and detailed modeling of author trustworthiness. In future, we plan to further explore stronger sentiment analysis to get better scoring functions. It would also be interesting to study bias in opinions expressed in forums, in case some forums talk primarily in support of alternate medicines and oppose other treatment options. In the forums we studied, we did not find specific biases towards or against a treatment, and hence this was not a factor in this work. Further study is also required to study the effects of spamming. Though we preprocessed the corpus to remove duplicate text from posts, stronger spam identification and removal may help reduce spurious results.

## 8. ACKNOWLEDGMENTS

We would like to thank Prof. Bruce Schatz from the University of Illinois for sharing data and domain-specific insights. The survey results reported in this paper were obtained from searches of the iPOLL Databank provided by the Roper Center for Public Opinion Research[4], University of Connecticut. This research was supported by the Multimodal Information Access and Synthesis Center at UIUC, part of CCICADA, a DHS Science and Technology Center of Excellence and the Army Research Laboratory under agreement W911NF-09-2-0053. Any opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily reflect the view of the sponsors.

## 9. REFERENCES

[1] Centers for Disease Control and Prevention. `http://www.cdc.gov/`.

[2] Food and Drug Administration. `http://www.fda.gov/`.

[3] GuidedMed. `http://www.guidedmed.com/en/`.

[4] Health on Net Foundation. `http://www.hon.ch/`.

[5] MeSH Ontology. U.S. National Library of Medicine. `http://www.nlm.nih.gov/mesh/`.

[6] Survey by Harris Interactive, July 7-July 14, 2009.

[7] Survey by Henry J. Kaiser Family Foundation, January 26-March 8, 2009.

[8] The Lemur Project. `http://www.lemurproject.org/`.

[9] WebMD®. `http://www.webmd.com/`.

[10] Wikipedia. `http://en.wikipedia.org/`.

[11] Yahoo! health portal. `http://health.yahoo.com/`.

[12] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE TKDE*, 17(6):734–749, 2005.

[13] M. Banko, M. Cafarella, M. Soderland, M. Broadhead, and O. Etzioni. Open Information Extraction from the Web. In *Proc. of IJCAI*, pages 2670–2676, 2007.

[14] P. F. Brown, V. J. D. Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479, 1992.

[15] M.-W. Chang, Q. Do, and D. Roth. Multilingual Dependency Parsing: A Pipeline Approach. In *Recent Advances in NLP*, pages 55–78, 2006.

[16] B. Chee, R. Berlin, and B. Schatz. Measuring Population Health using Personal Health Messages. *AMIA 2009 Annual Symposium*, 2009.

[17] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[18] A. Gaudinat, P. Ruch, M. Joubert, P. Uziel, A. Strauss, M. Thonnet, R. Baud, S. Spahni, P. Weber, J. Bonal, C. Boyer, M. Fieschi, and A. Geissbuhler. Health Search Engine with e-Document Analysis for Reliable Search Results. *Intl. Journal of Medical Informatics*, 75:73–85, 2006.

[19] A. Kittur and R. E. Kraut. Harnessing the Wisdom of Crowds in Wikipedia: Quality through Coordination. In *Proc. of CSCW*, pages 37–46, 2008.

[20] D. Lin. Automatic Retrieval and Clustering of similar words. In *COLING-ACL*, pages 768–774, 1998.

[21] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[22] M. Paşca. Organizing and Searching the World Wide Web of Facts - Step Two: Harnessing the Wisdom of the Crowds. In *Proc. of WWW*, pages 101–110, 2007.

[23] M. Paşca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. Organizing and Searching the World Wide Web of Facts - Step One: the One-Million Fact Extraction Challenge. In *Proc. of AAAI*, pages 1400–1405, 2006.

[24] D. Roth, M. Sammons, and V. Vydiswaran. A Framework for Entailed Relation Recognition. In *Proc. of $47^{th}$ Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 57–60, 2009.

[25] V. Vydiswaran, C. Zhai, and D. Roth. Content-driven Trust Propagation Framework. In *Proceedings of the $17^{th}$ SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.

[26] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proc. of EMNLP*, pages 347–354, 2005.

[27] X. Yin, J. Han, and P. S. Yu. Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008.

[28] T. Zesch and I. Gurevych. Wisdom of Crowds versus Wisdom of Linguists – Measuring the Semantic Relatedness of Words. *Natural Language Engineering*, 16(1):25–59, 2009.

[4] `http://www.ropercenter.uconn.edu/`