AIAA Guidance, Navigation, and Control Conference
08 - 11 August 2011, Portland, Oregon

AIAA 2011-6528

A Method for Using Historical Ground Delay Programs to Inform Day-of-Operations Programs

Shawn R. Wolfe[*] and Joseph L. Rios[†]

NASA Ames Research Center, Moffett Field, CA, 94035

While each day in the National Airspace System is unique, even the most challenging traffic flow management situations are likely to have some historical precedent. Unfortunately, traffic managers today are mostly limited to their own past experiences, and cannot leverage the experiences of others. This paper describes a novel approach for leveraging such "lessons learned" by making past control actions in similar situations available to the decision maker, specifically in the context of potential ground delay programs. By reviewing the response in similar situations, the decision maker is better informed as to whether a ground delay program would be needed, and if so, what the duration, scope, and rates could be. In the work presented here, methods for searching a combined historical archive of weather, traffic and operational actions for comparable conditions are evaluated. Multiple Weather Impacted Traffic Indices, describing both observations and forecasts, are used to characterize the operational situation. The control action response is extracted from the National Traffic Management Log and posted advisories. We evaluate several schemes for finding relevant past situations by their ability to retrieve situations with the same control action as was ultimately chosen by the traffic managers. All but one scheme exceeded baseline performance in each retrieval category, with an average improvement of 10% on the range of the metrics we used, suggesting that an appropriate decision support tool could inform decisions on ground delay programs.

Nomenclature

FAA	Federal Aviation Administration
GDP	Ground Delay Program
IMC	Instrument Meteorological Conditions
MAP	Mean Average Precision
METAR	Aviation Routine Weather Report
MRR	Mean Reciprocal Rank
NASA	National Aeronautics and Space Administration
OEP	Operational Evolution Partnership
WITI	Weather Impacted Traffic Index
S	Set of vectors
Q	Set of query vectors
R	Set of relevant vectors
X	Set of candidate vectors
$f()$	Arbitrary function

I. Introduction

Air traffic controllers work to maintain safe operations by directing aircraft departing from and arriving at airports. However, degradation of operating conditions or an increase in traffic (or both) at an

[*] Computer Scientist, Intelligent Systems Division, MS 269-2, Shawn.R.Wolfe@nasa.gov. Member AIAA.

[†] Aerospace Engineer, Systems Modeling and Optimization Branch, MS 210-15, Joseph.L.Rios@nasa.gov. Member AIAA.

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

airport can create difficult situations for the controllers. In order to prevent this from occurring, traffic flow managers monitor the system and issue traffic management initiatives to keep traffic at manageable levels. Ground delay programs are among the most aggressive traffic management initiatives, but if implemented at the wrong time or with the wrong parameters, may also decrease the efficiency of the National Airspace System. A ground delay program assigns arrival slots to a set of traffic destined for a particular airport, often resulting in a variable amount of pre-departure delay for each affected flight. Though pre-departure delay (holding on the ground) is generally preferred to airborne delay (holding in the air), it may result in unnecessary delay should conditions improve at the destination airport sooner than expected.

Today's traffic flow managers rely on personal experience to create traffic management plans, including the issuance of ground delay programs. This has some potential drawbacks. First, two equally experienced managers may create equally effective but different plans for the same situation. Such unpredictability in response creates uncertainty for the airspace users, who as a result may fail to take proactive actions because of this variability. Second, less experienced traffic flow managers might develop less optimal plans: either less effective, too severe, or both inadequate and overly restrictive in some aspect. Finally, since no traffic flow manager has been on hand for all prior conditions, an unfamiliar situation may arise that would be challenging to effectively and efficiently manage without some decision support.

We propose to address these issues by developing a decision support system that provides access to historical traffic flow management initiatives. These initiatives will be indexed by the relevant traffic and weather conditions that led to the actions taken by the traffic flow manager. Such a system would enable a traffic flow manager to expand upon personal knowledge by including the perspective of other traffic flow managers through a review of their past actions. At the core of this system is the ability to search for actions under a set of stated weather and traffic conditions, which may either be the current conditions or a "what-if" scenario. For example, the traffic flow manager could provide a traffic pattern and weather phenomenon to the decision support tool and inspect the actions taken in ten similar circumstances. This would suggest what actions could be taken, as well as which ones were taken more often, and should lead to more predictable, efficient, and timely decision making. Crucial to the success of such a decision support system is its search algorithm, in particular how well it ranks historical situations in order to their similarity to the scenario query. What makes a set of conditions similar? In this paper, we evaluate several techniques for rating the similarity of set of operating conditions.

The remainder of the paper is organized as follows. A brief review of related work is provided in Section II. Sections III and IV describe the data used and metrics used in our study. Next, Section V outlines the models used to rank situations according to their estimated similarity to the scenario query. The results of our experiment is detailed in Section VI. Finally, we conclude with a summary and future work in Section VII.

II. Related Work

Researchers have previously explored search-based system for other weather-impacted fields, as well as other forms of traffic flow management decision support and prediction. Ji, Yuan and Yue used a case-based reasoning approach to retrieve representative weather cases from a database of weather events.[23] They used similarity measure on several weather attributes that is equivalent to our weighted sum form with a log transform. Jain, Srinivas and Rauta developed a fuzzy logic system to forecast power (energy) loads, based on weather conditions.[24] They used a similarity measure based on Euclidean distance, defined on several weather variables, to find the set of similar weather conditions to make the forecast. Juell and Paulson used weighted sum algorithm, as well as a neural net, to predict the dew point from other weather variables in a another case-based system.[25] Reinforcement learning was used to tune the values of the weights. Elmore and Richman also use the Euclidean distance to find modes in weather data.[26] They relate their choice of similarity metric to the more general Minkowski distance, and suggest other forms (such as the weighted sum) also be considered. Finally, Klein used weather-impacted traffic indices, as we do, to characterize weather in the National Airspace System.[27] He used the cosine model to find similar days in terms of conditions at selected airports, but as he also factored in the overall magnitude, the approach is more comparable to the Euclidean distance model.

To our knowledge, the proposed ability to search for past traffic management initiatives by the similarity of operating conditions is unique. The original concept of the decision support tool was suggested by Rios.[28] The most analogous decision support tools, such as the Enhanced Traffic Management System,[29] provide situational awareness but do not provide historical context. Smith, Sherry and Donahue investigated the possibility of predicting GDPs.[30] They used a support vector machine, trained on historical weather forecasts and airport arrival rates, to predict an arrival rate and GDP status given a new weather forecast. Nonetheless, the envisioned usage is different from our proposed decision support tool, as they recommend a particular action, whereas we propose to provide historical context. Our study fills the gap between decision support tools that only provide information on the current situation (and not access to past decisions), and prediction tools which utilize historical data but do not support the user in making decisions from that historical perspective.

III. Dataset

A search system that meets our specifications would need to include both the situational parameters considered by the decision makers, as well their ultimate decisions. We fused two datasets, described below, to create our representation of the situation and corresponding action. These datasets provided hourly data from 2008 and 2009 for the FAA's OEP-35 (with the exception of Honolulu), which consists of the busiest U.S. airports. From this set, we eliminated any airport that did not have a minimum of fifty ground delay programs (GDPs) issued in 2009, leaving the airports listed in table 1.

Table 1. Weather-related GDP Statistics for airports in our dataset

IATA Code	Airport	GDPs in 2009	Total hours of GDP in 2009
ATL	Hartsfield-Jackson Atlanta International	213	824
BOS	Boston Logan International	91	523
CLT	Charlotte Douglas International	77	191
DEN	Denver International	67	162
EWR	Newark Liberty International	227	1959
JFK	New York John F. Kennedy International	152	852
LGA	New York LaGuardia	146	1406
ORD	Chicago O'Hare International	90	503
PHL	Philadelphia International	142	857
SFO	San Francisco International	195	1006

A. Weather Impacted Traffic Index

Both weather and traffic must be captured in our representation of the relevant conditions. Both are rich and complex, and their effective representation is an area of research in itself. Fortunately, we can leverage previous research that lead to the development of the Weather Impacted Traffic Index (WITI)[1] to represent both weather and traffic. WITI is a model which seeks to capture specifically the effect weather has on traffic. We use a version of WITI that includes a forecast component that had been previously developed by other researchers.[2] This forecast is important to consider in our search system, as GDPs are planned not only based on the current conditions, but also on what is expected in the future. The weather forecasts were obtained for two hours, four hours, and six hours in the future. The traffic forecast was generated from the scheduled traffic. The weather forecasts came from the Collaborative Convective Forecast Product[3] and Terminal Aerodrome Forecasts,[4] whereas current weather conditions came from the National Convective Weather Diagnostic[5] and Aviation Routine Weather Reports (often referred to by the French acronym METARs).[6]

We use a form of WITI that is broken down into seven separate WITIs that capture different impacts at a given airport. These are:

1. En-route convective weather, which captures the impact of convective weather on incoming and outgoing traffic, up to five hundred nautical miles away.

2. Local convective weather, which captures the impact of convective weather within one hundred nautical miles of the airport.

3. Wind, with or without accompanying precipitation.

4. Snow, which captures all sorts of cold hazards.

5. Instrument Meteorological Conditions (IMC), which captures the impact of poor visibility on traffic.

6. Traffic volume, which captures volume impacts caused by weather disruptions elsewhere in the system.

7. Other, capturing miscellaneous weather-related impacts that do not fall into the categories above.

As we have the current conditions in addition to three forecast periods for each of the seven WITIs, we have 28 WITIs to characterize the conditions at each hour at any of the previously listed airports. Each WITI is a nonnegative real number. Fig. 1 shows an example of WITI scores where delays due to en-route convection and wind are predicted to increase, according to the forecasts.

Figure 1. Example of WITI representation (from JFK airport).

	en-route convection	local convection	wind	snow	IMC	volume	other
Now	12.2	0.0	0.0	0.0	0.0	0.0	0.2
2-hour forecast	64.7	0.0	9.2	0.0	0.0	0.0	0.0
4-hour forecast	56.3	0.0	111.4	0.0	0.0	0.0	0.0
6-hour forecast	168.8	0.0	70.8	0.0	0.0	0.0	0.0

B. The National Traffic Management Log and Advisories Database

Traffic flow management decisions are made at various centers across the U.S. Fortunately, the difficult task of recording, standardizing and centralizing these decisions has already been done. Originally designed to provide situational awareness, the National Traffic Management Log (NTML)[7] also contains an archive of traffic flow management decisions in a database. This database contains both initial decisions as well as any subsequent revisions. An earlier study shows that the unified NTML has provided numerous benefits over the prior, loosely controlled logging system.[8] We use the historical archive of the NTML to provide an hourly snapshot for the GDP status of the airports in our study.

In addition to the NTML, we also scraped pages from the FAA's online advisory database.[9] We matched these records to those in the NTML, cross-checking the data and also adding information about when the advisory was publicized.

C. Processing

We combined our WITI and NTML/advisory data to give us 8688 instances in 2008 and 8544 instances in 2009 of current and forecast conditions (from the WITI data source) and corresponding actions (from the NTML/advisory data sources), as given in figure 2. The 28 WITIs are as described in Section A (though the current conditions and three forecasts are not separately represented in figure 2). We treated ground stops as a type of GDP. The GDP cause is the reason given for the GDP, or "no GDP" if no GDP is active. However, over half of the GDP cause types in our database had no obvious relationship to weather (e.g., "Air Show"). We treated all such non-weather GDPs as "no GDP" in our dataset. Likewise, some GDP causes were very similar from the WITI perspective (e.g., "Fog" and "Low Visibility"), and were mapped into a joint cause. In all, this left us with four GDP causes: in order of prevalence, they are "wind", "visibility", "convection", and "snow", and "no GDP". The scheduled start and end time reflects the proposal at that particular time and not the actual start and end times (the times a GDP is planned to occur are not always the same as when the GDP actually occurs); furthermore, these times are not defined when the cause is "no GDP". For this study, we did not track any other parameters of the GDP, such as scope and rate.

WITI has been found to have a high correlation to overall national delays[10] and delays at specific airports.[11] However, this does not necessarily mean that WITI values are strongly indicitive of ground delay programs. We ran a simple experiment to evaluate the relationship between our WITI scores and ground

Figure 2. Instance representation for weather conditions and corresponding GDP status

en-route convection	local convection	wind	snow	IMC	volume	other	GDP cause	scheduled start time	scheduled end time

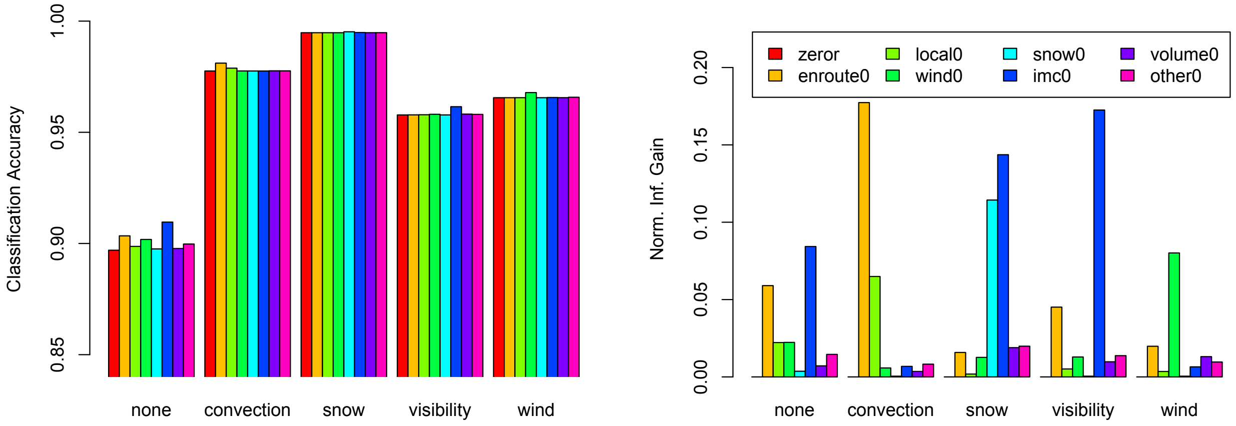

Figure 3. Correspondence of each boolean WITI to the GDP cause categories for all airports in 2008, in terms of accuracy and normalized information gain. Only the current conditions were compared against the current GDP state.

delay program causes (including "no GDP", as described above). We cast this as a classification problem to determine if a ground delay program is in effect, further divided into five separate classification problems, one for each of the five causes. For each WITI type, we found the rule with the highest classification accuracy by retrospectively picking the best WITI cutoff threshold (with different thresholds for different causes and airports). Each rule would make its prediction by comparing the actual WITI value with this cutoff threshold. For instance, a rule could be "predict a GDP due to snow in Atlanta if the observed snow WITI score exceeds 200." As the observed (non-forecast) WITIs are presumably more accurate than the forecast WITIs, we only evaluated the accuracy of these WITIs against the current GDP status (instead of planned GDPs, for the same reason). For further context, we compare these results to a baseline constant rule dubbed "zeror" (meaning zero rule), which always gives the same prediction without using any data.

Figure 3 shows our evaluation on two measures, classification accuracy and normalized information gain over all airports, for the 2008 data. The leftmost chart shows classification accuracy, which is the fraction of times a prediction based on the WITI would yield the correct answer. All seven WITIs yield highly accurate predictions, but so does the constant "zeror" rule. This is because GDPs are somewhat uncommon, and always guessing that there is not a GDP produces highly accurate results. It is difficult to discern how much is gained by using the individual WITIs when evaluating classification accuracy, so a different evaluation measure is needed. The rightmost chart shows how much information can be gained by considering the individual WITI, according to information theory.[12] We normalized this information gain to be on a 0-1 scale, so that 0 is no information and 1 is complete information. Since the "zeror" rule makes the same prediction in all cases, it has no information gain, and therefore cannot be seen on the chart. Unfortunately, the individual WITIs do not fare much better, with none providing even a fairly modest 20% gain in information. This means that searching for similiar events by WITI will not necessarily lead to events with the same control action, and that producing good search results will be challenging.

IV. Evaluation Metrics

Core to the evaluation of information retrieval results is the notion of relevancy. In a typical information retrieval setting, the items in the search collection (known as a corpus) are evaluated for relevancy against several queries by either the users themselves or trained assessors. Typically, relevancy is seen as a binary measure (relevant/non-relevant), though levels of relevancy are sometimes used.

This study needed to overcome the lack of available relevance data from actual traffic flow managers. However, since we know both the task (to make a decision regarding a possible GDP) and what decision was ultimately made, we can estimate what should be relevant. We cast the problem as that of task-based information retrieval,[13,14] and use our understanding of the task and decision to simulate queries and relevance judgments. Specifically, as we regard the task as making a control action decision, a reasonable query is the observed and forecast conditions for that time. The query consists of only the WITIs (see figure 2, above) and not the GDP status, as the latter is the decision to be made by the user of the system. The historical instances from the same airport are used as the corpus.

Once again, we use the envisioned usage of the system to simulate the relevance judgments. Since the decision support system is designed to assist traffic flow managers in making decisions on GDPs, we use the GDP status to derive the relevancy judgments. We treat the GDP decision that was made by the traffic flow managers in the situation as the (only) correct choice. Presumably, a system that favors instances from the corpus with the same GDP decision as was chosen for the queried situation would have been helpful for making that decision. Therefore, we define relevancy as a match between the corresponding action of the retrieved instance and the one chosen in the queried situation. (Note that the GDP actions are used only to define relevance, and are not a part of the query or similarity calculations.)

However, in addition to the cause, the GDP decisions also have scheduled start and end times, which present a challenge to map to binary relevance. We opted to reformulate the problem as several parallel problems, each of the form "Will we need a GDP Y hours from now?", where Y was specified in hour increments up to the 6-hour WITI forecast, for a total of seven parallel problems. In practice, some instances with a different GDP status would be seen as similar, while some with the same might not (as conditions can vary greatly among situations with the same GDP status), but we believe our approximation to be as reasonable as possible without human assessors and without biasing the results by using WITIs in the relevance definition.

Given our simulated queries and relevance judgments, we can evaluate the ranking of instances (as described in Section V) provided by the search system. The metrics we use are based on a metric defined on unranked results called *precision*. Given set of results produced for a query, the precision is simply the fraction of the set that is relevant to the query. Formally, given a query q, a set of results S_q and the set of all relevant items R_q, the precision is given in Eq. (1).

$$p(S_q, R_q) = \frac{|S_q \cap R_q|}{|S_q|} \tag{1}$$

To translate this to a ranked list of results, we define a series of overlapping sets by including only the top m ranked results. We choose m as the smallest number that includes a specified number of relevant results, so that $S_{q,1}$ includes a single relevant result, $S_{q,2}$ includes two relevant results, and so on up to $S_{q,|R_q|}$.

One metric is the precision of $S_{q,1}$, which is smallest top m-ranked search results that include a single relevant result. This metric is known as the *reciprocal rank*, as it is equivalent to the reciprocal of the rank of the highest ranked relevant instance. It corresponds to the number of search results a user would need to inspect in order to find a single relevant result, when evaluating the search results in ranked order. It is a good metric when the user's task can be satisfied by any relevant result. Since we have a set of queries, we use the *mean reciprocal rank* (MRR)[15] as one of the two metrics we use to evaluate our results. Given the set of queries Q and other quantities as defined for Eq. (1), the mean reciprocal rank is defined in Eq. 2.

$$MRR(Q) = \frac{1}{|Q|}\sum_{q\in Q} p(S_{q,1}, R_{q,1}) = \frac{1}{|Q|}\sum_{q\in Q}\frac{|S_{q,1}\cap R_q|}{|S_{q,1}|} = \frac{1}{|Q|}\sum_{q\in Q}\frac{1}{|S_{q,1}|} \tag{2}$$

However, there are shortcomings of the MRR metric from the aspect of our evaluation, stemming from the reciprocal rank on which it is based. Reciprocal rank models a user need that is satisfied with a single relevant result. However, it is not obvious that this would be the case for our decision problem, in fact it seems more likely that the traffic flow manager would need to review several comparable situations to arrive at a decision. Also, because the reciprocal rank is based on the rank of a single item, it tends to be a metric with high variance, though this is reduced by the averaging over queries in that occurs in the MRR metric.

Instead of reciprocal rank, the *average precision* can be used, which is precision averaged over all sets $S_{q,n}$ (where reciprocal rank used only $S_{q,1}$). Once again, this can also be averaged over a set of queries, leading to the *mean average precision* (MAP),[16] given in Eq. 3.

$$MAP(Q) = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{|R_q|} \sum_{i=1}^{|R_q|} p(S_{q,i}, R_{q,i}) = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{|R_q|} \sum_{i=1}^{|R_q|} \frac{i}{|S_{q,i}|} \tag{3}$$

The MAP metric does not assume a certain number of relevant results are needed for the user's task, instead averaging over all possible levels. This better fits our uncertainty of the traffic flow managers' needs, and is more commonly used in information retrieval evaluations.[17]

Finally, as we have different categories of results (both from different airports and from queries of different GDP causes), we follow the example of Sebastiani[18] and use the *microaverage* to combine averages from different categories. Given an evaluation metric $f()$ and k sets $S_1, \ldots, S_k$, the microaverage over $f()$ is given in Eq. 4.

$$f_{micro}(S_1, \ldots, S_k) = \sum_{i=1}^{k} \frac{|S_i|}{\sum_{j=1}^{k} |S_j|} f(S_i) \tag{4}$$

In our use, $f()$ is either MRR or MAP, and since these are both averages themselves, the microaverage simplifies to the corresponding metric over all queries without regard to the category.

V. Models

We experimented with standard models from the information retrieval, machine learning and operations research communities to produce our rankings. We use a utility-based approach for ranking, which means that each instance is given an ordinal score (or utility, in this case a similarity measure), and the ranking is defined by the corresponding order of the scores. In the following, the query Q and instance X consist of the 28 WITIs (see Section C and fig. 2), with q_i and x_i denoting the i^{th} WITI value of the query and instance, respectively.

Perhaps the most widely used similarity measure in information retrieval is the durable cosine,[19] defined in our domain as in Eq. 5.

$$cos(Q, X) = \frac{\sum_{i=1}^{28} q_i x_i}{\|Q\| \, \|X\|} \tag{5}$$

Since the WITIs are all nonnegative, the cosine will range from 0 to 1, with higher scores indicating more similarity and thus higher ranked. An important feature of the cosine similarity is that it is independent of the overall magnitude of both the query and instance vectors, a feature that makes sense for document retrieval but does not appear to be a good match for our domain. Rather, the cosine measure is sensitive only to the relative distribution among the vectors of the query and instance.

Likewise, the common Euclidean distance is often used in machine learning algorithms,[20] defined on an multidimensional space. Since we have 28 WITIs, we have 28 dimensions, and so the Euclidean distance is defined as in Eq. 6.

$$distance(Q, X) = \sqrt{\sum_{i=1}^{28} |q_i - x_i|^2} \tag{6}$$

Since Eq. 6 is a distance measure, larger values indicate less similarity and thus lower ranked. A feature of the Euclidean distance is the possibility of a "shortcut", i.e., that dissimilarity is lessened when the differences are distributed across different dimensions (compared to an equal total magnitude in the same dimension). It is unclear if this applies to our domain.

Finally, we employ a simple weighted sum from operations research,[21] in an even simpler form as all our weights are 1. With equal weighting used, the weighted sum we used in our experiments is given in Eq. 7.

$$wsum(Q, X) = \sum_{i=1}^{28} |q_i - x_i| \tag{7}$$

Like the Euclidean distance, the value of Eq. 7 grows as the differences increase, and we rank in decreasing order of scores. Implied by its formula, the weighted sum model assumes that the same differential change on a particular dimension will always result in the same change in the score, regardless of the other dimensions. Though it would stand to reason that any increase in WITIs should correspond to worse conditions, this independence property may not hold in our domain.

Though the three models described have different properties and produce different rankings, they are in fact all equivalent to the parameterized Minkowski distance measure,[22] with an additional modification for the cosine model. The Minkowski distance takes an additional parameter p and is defined in our domain as in Eq. 8.

$$Minkowski(Q, X, p) = \left(\sum_{i=1}^{28} |q_i - x_i|^p \right)^{1/p} \tag{8}$$

The relationship to the Euclidean distance and the (evenly) weighted sum should be obvious by comparing the formulas. As it turns out, the ranking produced by the cosine model is equivalent to that of the Euclidean distance model if the vectors are first normalized to unit vectors. This may be easier to see when considering the geometry of the problem.

As an alternative to the WITI values as given, we also experimented with a transformation of the WITI scores. Each WITI appeared to be approximately distributed by an exponential distribution, with many WITI values falling in the low range, and fewer and fewer spread out on the long right tail. We created transformed WITIs by mapping these values onto an approximation of their empirical cumulative distribution function. This has some potential advantages. First, the cumulative distribution function ranges from 0 to 1, so all transformed WITIs are on the same scale. Second, differences in the long tail are minimized, matching the intuition that large WITI values may capture qualitatively the same impact (at least to the fidelity in which we model the GDP). On the other hand, small differences among the more frequent low WITI values will be seen as more meaningful.

The three original similarity models, combined with the same models using transformed WITIs, gives us six models to evaluate in experimentation. In addition to this, we added a random model to serve as a baseline. The random model scores instances randomly and produces a random ranking. Unlike our other models, a model that incorporates randomness can produce different results for different runs on the same data. To avoid this variability, we simply calculated what the theoretical average value of our metrics for the random model rather than report metrics on a limited number of runs.

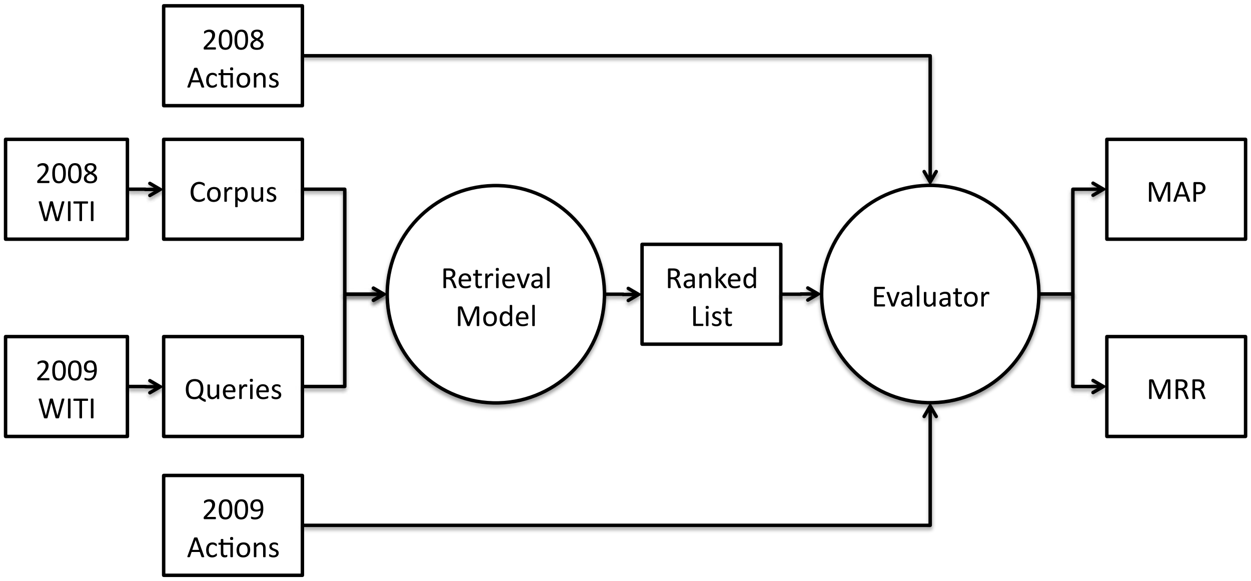

Figure 4. Flow of data into evaluation results in our experiment.

VI. Results

We used all the instances from 2009 as queries, to be matched against instances from 2008 corpus for the same airport. This simulates a decision support tool built on 2008 data that went "live" in 2009. For each query, all instances from the same airport in the corpus were scored by each of the models in Section V and ranked in a list. Our metrics were recorded separately for each GDP cause, as well as the microaveraged results, for each airport. The qualitative results vary from airport to airport and are too numerous to report. Instead, we use the microaverage (as given in Eq. 4) over GDP causes to combine the results from different airports in the following results. In addition to breaking down the results by GDP type, we also evaluate the performance on queries that we consider "hard". We define the query as "hard" if for the given time period, the decision was to change the plan from what was in place previously. This could be changing the start or end times for a GDP (including cancellations), or changing the GDP cause.

Figure 4 shows the data flow in our experiment. The WITI data (including the forecasted WITI, as in Fig. 1) is used as our representation of the operational situation. We divide this into two sets: the 2008 WITIs are used as the corpus (the body of data we will retrieve from), and the 2009 WITIs are used as our query set. Each query is evaluated by the retrieval models (described above in section V), with each producing a ranked list. The quality of these ranked lists are evaluated by comparing the corresponding actions of the query (from 2009) and each element in the ranked list. The final result is the MRR and MAP metrics, as defined in section IV and detailed below.

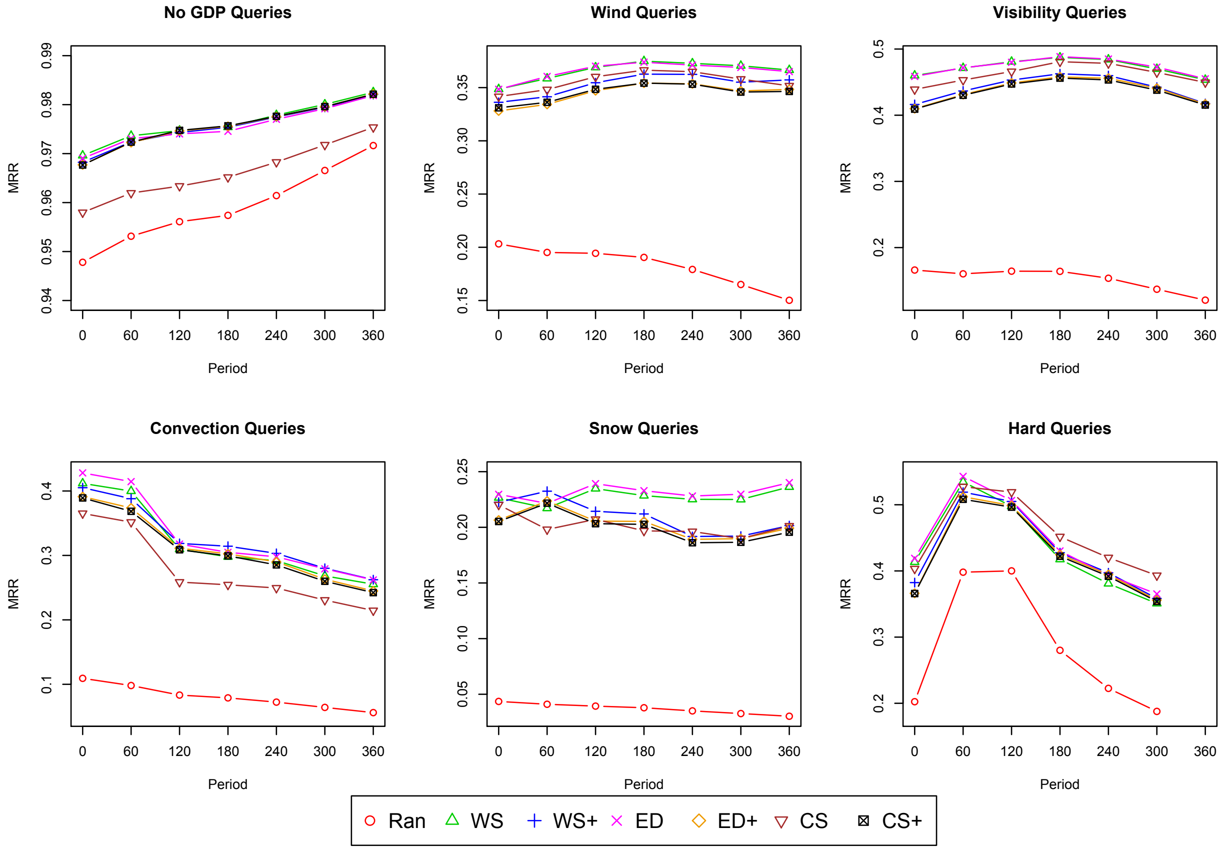

Figure 5. MRR results per GDP cause category, microaveraged over airport. WS is weighted sum, ED is Euclidean distance, CS is cosine, Ran is random, and '+' version indicate WITI transformation.

Figure 5 shows MRR results microaveraged over airport and separated by the GDP cause of the query. Except for the cosine model on queries with no GDPs, all models easily outperform the random model on the MRR metric. Retrieving an instance without a GDP should be easy, since most of the time no GDP is in place. Therefore, even though the evaluation measure is high, the fact that the cosine model only slightly outperforms the random model when there is no GDP is unsatisfactory. For the queries that do have a GDP, the performance of the non-random models is largely comparable, though the cosine and transformed WITI cosine underperform for certain GDP types. Among the remaining models, it is nearly impossible to have a preference.

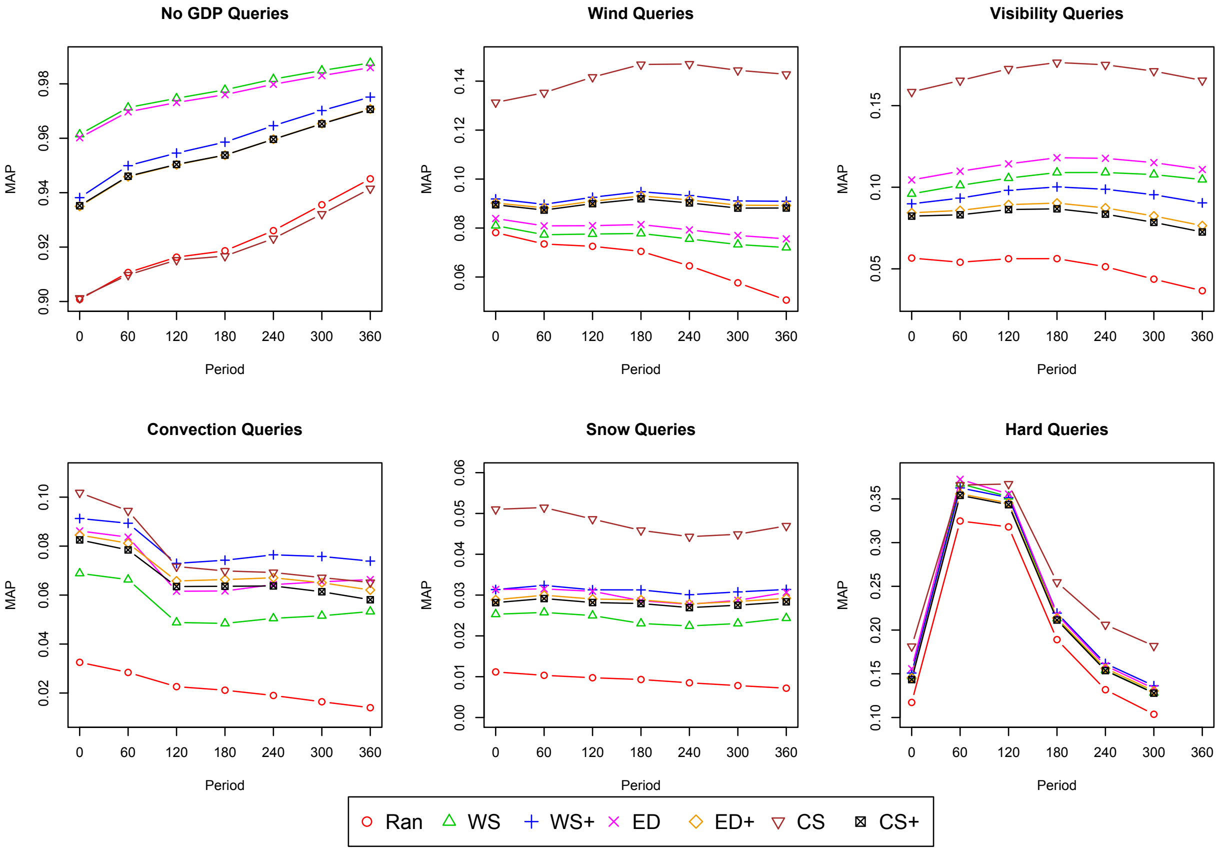

Figure 6. MAP results per GDP cause category, microaveraged over airport. WS is weighted sum, ED is Euclidean distance, CS is cosine, Ran is random, and '+' version indicate WITI transformation.

Figure 6 shows MAP results microaveraged over airport and separated by the GDP cause of the query. The differences between models is somewhat more pronounced, though it is still difficult to pick a best or worst model. Again, except for cosine, all models clearly outperform the random model. The cosine remains the puzzling member of the bunch; it's performance is as poor as random guessing when no GDP is in place, as with the MRR metric, and it has a mediocre showing on the convection queries. On the other hand, it clearly outperforms the other models on the visibility, wind, and snow queries. For the most part, the other models give roughly similar performance.

Initially, the performance of the cosine model, particularly when evaluated on mean average precision, was perplexing. The only difference between the cosine and Euclidean distance in terms of ranking is that cosine is not sensitive to the overall "length" (i.e., Euclidean norm) of the Weather Impacted Traffic Index scores, and the two give the same ranking when all instances have the same length. This would appear to be a disadvantage of the cosine model, and yet it has the best mean average precision performance for three of the four ground delay program causes.

As it turns out, this Weather Impacted Traffic Index length varies considerably among ground delay programs with the same cause. This means that some instances with a given ground delay program action will have WITI scores much closer to instances without a ground delay program action than those with large lengths with the same ground delay program cause. The only ground delay program cause that did not have such a large length variance were the convection ground delay programs, and not surprisingly the cosine model underperformed for convection. By the same token, the transformed Weather Impacted Traffic Index (see Section V) reduces the Weather Impacted Traffic Index length variance, which explains why the weighted sum and Euclidean distance performed better on retrieving ground delay programs with the transform in place. Overall, if we were to pick a single model from this set, we would use either the weighted sum or Euclidean distance with the Weather Impacted Traffic Index transform for this reason.

VII. Conclusions and Future Work

In this paper, we have proposed a decision support tool to assist traffic flow managers in deciding if a ground delay program is needed. The tool would assist not by making a recommendation, but by allowing the traffic flow manager to inspect past decisions by other traffic flow managers in similar conditions. We proposed six models to rank instances by similarity to given conditions and evaluated these on a database comprised of Weather Impacted Traffic Indices and corresponding ground delay program decisions for ten busy U.S. airports.

All six methods evaluated outperformed a random ranking model on the metrics of mean reciprocal rank and mean average precision, both validating the choice of Weather Impacted Traffic Index to represent conditions and showing that the proposed decision support tool could help guide users towards events with comparable control actions. What is not clear is which of the models would be the best choice. This is difficult to assess as the models have different relative performance on different query types, as particular models are apparently better at detecting some events and worse at detecting others. The query types were not evenly represented in our database, nor do we feel that each query type is equally important to traffic flow managers, though we do not how important each type might be.

In reality, none of the models we evaluated are likely to be the best for the task. One interesting possibility would be to evaluate additional parameter values for the Minkowski distance, or even derive it directly from the data. Another approach would be to combine the models we experimented with in some way. In particular, is there some way to get the advantages of the cosine models without incurring the cost? An alternative approach would be to use an automatic algorithm, such as boosting,[31] to combine the different models in a machine learning context.

Indeed, there are quite a few possibilities for learning from the data to improve performance. In this paper, we have already used the data to form the empirical exponential distribution for the Weather Impacted Traffic Index transformation. Other transformations could be explored, allowing for a more powerful fit of the data. Another possibility for learning would be to learn weights or linear rescalings of the Weather Impacted Traffic Indices. There are precedents for such with each approach, and an uneven weighting should be expected for our domain as Weather Impacted Traffic Indices are not equally accurate.

Finally, feedback from traffic flow managers would greatly improve the potential usefulness of the proposed decision support system. Indications of which instances are relevant to a query would give more accurate evaluations and allow us to better tune our models. Prioritization of the various ground delay program causes would allow us to create a more informed cost function to evaluate models. Lastly, additional input would refine our understanding of the needs of the traffic flow managers and better enable us to design a decision support tool to assist them.

Acknowledgments

The authors would like to thank Sasha Klein for providing the WITI data, Rich Jehlen for providing access to the NTML and formulation of the proposed decision support tool, and Shon Grabbe for input and support for this project.

References

[1] Callaham, M. B., DeArmon, J. S., Cooper, A., Goodfriend, J. H., Moch-Mooney, D., and Solomos, G., "Assessing NAS Performance: Normalizing for the Effects of Weather," *4th USA/Europe Air Traffic Management R&D Symposium*, 2001.

[2] Klein, A., Kavoussi, S., Hickman, D., Simenauer, D., Phaneuf, M., and MacPhail, T., "Predicting Weather Impact on Air Traffic," *Integrated Communications Navigation and Surveillance (ICNS) Conference*, 2007.

[3] National Weather Service, "Collaborative Convective Forecast Product Product Description Document," http://aviationweather.gov/products/ccfp/docs/pdd-ccfp.pdf.

[4] National Weather Service, *Weather Service Operations Manual*, chap. D-31, National Oceanic and Atmospheric Administration, 1997.

[5] National Weather Service, "National Convective Weather Forecast Product, Version 2 (NCWF-2)," http://products.weather.gov/PDD/ncwf-2_pdd.pdf.

[6] Office of the Federal Coordinator for Meterological Services and Supporting Research, *Federal Meteorological Handbook No. 1*, chap. 2, National Oceanic and Atmospheric Administration, 2005.

[7] Federal Aviation Administration, *Facility Operation and Administration*, chap. 17-5, U.S. Department of Transportation, 2010.

[8] Brickman, B. and Yuditsky, T., "Improving the Usability of an Automated Tool for the Recording, Coordination, and Communication of Traffic Managment Initiatives," *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, 2004, pp. 46–50.

[9] Federal Aviation Administration, "Air Traffic Control System Command Center Advisories Database," http://www.fly.faa.gov/adv/advAdvisoryForm.jsp.

[10] Chatterji, G. B. and Sridhar, B., "National Airspace System Delay Estimation Using Weather Weighted Traffic Counts," *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2005.

[11] Klein, A., Craun, C., and Lee, R. S., "Airport delay prediction using weather-impacted traffic index (WITI) model," *2010 IEEE/AIAA 29th Digital Avionics Systems Conference (DASC)*, 2010.

[12] Cover, T. M. and Thomas, J. A., *Elements of Information Theory*, chap. 2, Wiley, 2nd ed., 2006, pp. 13–52.

[13] Hersh, W., Pentecost, J., and Hickam, D., "A task-oriented approach to information retrieval evaluation," *Journal of the American Society for Information Science*, Vol. 47, 1996, pp. 50–56.

[14] Vakkari, P., "Task-based information searching," *Annual Review of Information Science and Technology*, Vol. 37, No. 1, 2003, pp. 413–464.

[15] Voorhees, E. M., "TREC-8 Question Answering Track Report," *Proceedings of the 8th Text REtrieval Conference*, 1999.

[16] Voorhees, E. M. and Harman, D., editors, *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, 2005.

[17] Manning, C. D., Raghavan, P., and Schütze, H., *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[18] Sebastiani, F., "Machine learning in automated text categorization," *ACM Computing Surveys*, Vol. 34, No. 1, 2002.

[19] Salton, G., *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison Wesley, Reading, MA, 1989.

[20] Aha, D. W., Kibler, D., and Albert, M. K., "Instance-Based Learning Algorithms," *Machine Learning*, Vol. 6, No. 1, 1991, pp. 37–66.

[21] Triantaphyllou, E., *Multi-Criteria Decision Making Methods: A Comparative Study*, chap. 2, Kluwer Academic Publishers, 2000.

[22] Cox, T. F. and Cox, M. A. A., *Multidimensional Scaling*, chap. 3, Chapman and Hall/CRC, 2nd ed., 2000, p. 63.

[23] Ji, S., Yuan, S., and Yue, J., "A Method of Weather Cases Generation Based on Similarity Rough Set," *Proceedings of the International Conference on Management and Service Science (MASS '09)*, IEEE, 2009, pp. 1–4.

[24] Jain, A., Srinivas, E., and Rauta, R., "Short term load forecasting using fuzzy adaptive inference and similarity," *Proceedings of the World Congress on Nature & Biologically Inspired Computing (NaBIC 2009)*, IEEE, 2009, pp. 1743 – 1748.

[25] Juell, P. and Paulson, P., "Using Reinforcement Learning for Similarity Assessment in Case-Based Systems," *IEEE Intelligent Systems*, Vol. 18, No. 4, 2003, pp. 60 – 67.

[26] Elmore, K. L. and Richman, M. B., "Euclidean Distance as a Similarity Metric for Principal Component Analysis," *Monthly Weather Review*, Vol. 129, No. 3, 2001, pp. 540 – 549.

[27] Klein, A., "Day's Weather in the NAS: Visualization of Impact, Quantification, and Comparative Analysis," *Integrated Communications Navigation and Surveillance (ICNS) Conference*, 2006.

[28] Rios, J. L., "Aggregate Statistics of National Traffic Management Initiatives," *Proceedings of the 10th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, 2010.

[29] Federal Aviation Administration, "Enhanced Traffic Management System," http://hf.tc.faa.gov/projects/etms.htm.

[30] Smith, D. A., Sherry, L., and Donohue, G., "Decision Support Tool for Predicting Aircraft Arrival Rates, Ground Delay Programs, and Airport Delays from Weather Forecasts," *Proceedings International Conference on Research in Air Transportation (ICRAT-2008)*, 2008.

[31] Schapire, R. E. and Singer, Y., "Improved Boosting Algorithms Using Confidence-Rated Predictors," *Machine Learning*, Vol. 37, No. 3, 1999.